Πανεπιστήμιο Πειραιώς
Τμήμα Πληροφορικής

Ζητήματα Ομοιότητας στην Εξόρυξη Γνώσης - Μεθοδολογίες και Τεχνικές

Διδακτορική Διατριβή

Ειρήνη Χρ. Ντούτση
MSc, Διπλωμ. Μηχανικός Υπολογιστών & Πληροφορικής

Πειραιάς, Ιούνιος 2008

Εγκρίθηκε από την Εξεταστική Επιτροπή, ———- 2008

.....	.....	.....
Ι. Θεοδωρίδης	Μ. Βαζιργιάννης	Γ. Τσιχριντζής
Επίκ. Καθηγητής	Αναπ. Καθηγητής	Αναπ. Καθηγητής
Παν. Πειραιώς	ΟΠΑ	Παν. Πειραιώς
Επιβλέπων	Επιβλέπουσα επιτροπή	Επιβλέπουσα επιτροπή

.....	.....	.....
Γ. Βασιλακόπουλος	Δ. Δεσπότης	Τ. Σελλής
Καθηγητής	Καθηγητής	Καθηγητής
Παν. Πειραιώς	Παν. Πειραιώς	ΕΜΠ
Εξεταστική επιτροπή	Εξεταστική επιτροπή	Εξεταστική επιτροπή

.....

Ι. Σίσκος

Καθηγητής
Παν. Πειραιώς
Εξεταστική επιτροπή

Πρόλογος

Στις μέρες μας εξάγονται όλο και περισσότερα πρότυπα εξαιτίας της πληθώρας των δεδομένων και της ευρείας χρήσης της Διαδικασίας Ανακάλυψης Γνώσης από τα Δεδομένα και της Εξόρυξης Γνώσης. Αυτή η πληθώρα των προτύπων επιβάλει την αντιμετώπιση προβλημάτων που σχετίζονται με τη διαχείρισή τους. Μία από τις πιο σημαντικές λειτουργίες στα πρότυπα είναι αυτή της αποτίμησης της ομοιότητας μεταξύ προτύπων, ένα πρόβλημα που έχει πολλές εφαρμογές και εγείρει σημαντικά ερευνητικά θέματα.

Στα πλαίσια της παρούσας διατριβής μελετάμε διάφορα θέματα που προκύπτουν κατά την αποτίμηση της ομοιότητας μεταξύ προτύπων. Πιο συγκεκριμένα, προτείνουμε αρχικά ένα πλαίσιο για την αποτίμηση της ομοιότητας μεταξύ προτύπων αυθαίρετης πολυπλοκότητας τα οποία ορίζονται τόσο πάνω σε πρωτογενή δεδομένα όσο και πάνω σε άλλα πρότυπα. Στη συνέχεια μελετάμε προβλήματα ομοιότητας για μερικούς από τους πιο δημοφιλείς τύπους προτύπων, συγκεκριμένα για συχνά στοιχειοσύνολα, δέντρα απόφασης και συστάδες. Ειδικότερα για την περίπτωση των συχνών στοιχειοσυνόλων, εξετάζουμε κατά πόσο οι παράμετροι της Εξόρυξης Γνώσης επηρεάζουν το αποτέλεσμα της σύγκρισης μεταξύ συνόλων από στοιχειοσύνολα. Στην περίπτωση των δέντρων απόφασης, προτείνουμε ένα πλαίσιο που βασιζόμενο στα δέντρα απόφασης αποτιμά την ομοιότητα τόσο μεταξύ δέντρων απόφασης όσο και μεταξύ συνόλων δεδομένων κατηγοριοποίησης. Τέλος, στην περίπτωση των συστάδων προτείνουμε μέτρα απόστασης μεταξύ συστάδων και συσταδοποιήσεων, τα οποία και χρησιμοποιούμε στη συνέχεια για να παρακολουθήσουμε την εξέλιξη και να εντοπίσουμε τυχόν αλλαγές σε δυναμικούς πληθυσμούς. Κοινό στοιχείο σε όλες τις παραπάνω περιπτώσεις είναι η θεώρηση ότι τα πρότυπα αποτελούνται από μία δομική και μία ποσοτική συνιστώσα, γεγονός που ανοίγει νέους δρόμους προς την κατεύθυνση ενός ενιαίου μοντέλου για τα αποτελέσματα της Εξόρυξης Γνώσης.

Ειρήνη Ντούτση

Ευχαριστίες

Κουράγιο χρειάζεται. Ανάμεσα στο δείκτη του χεριού σου και την άκρη του τετραδίου σου απλώνεται τεραστίου μήκους έκταση που έχεις να διανύσεις.

Εκ του πλησίον, Οδυσσέας Ελύτης

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντά μου, Επικ. Καθ. Ιωάννη Θεοδωρίδη για την υποστήριξη και την καθοδήγησή του όλα αυτά τα χρόνια. Οι γνώσεις του και η εμπειρία του με βοηθήσαν όλες εκείνες τις φορές που τα προβλήματα φαινόντουσαν ανυπέρβλητα, και η στάση του με ενέπνευσε τόσο σε ερευνητικό όσο και σε ανθρώπινο επίπεδο.

Μεγάλο μέρος αυτής της δουλειάς είναι αποτέλεσμα συνεργασίας. Με την Καθ. Μάιρα Σπηλιοπούλου συνεργαστήκαμε εκτενώς σε θέματα εξέλιξης συστάδων και την ευχαριστώ ιδιαίτερα για τις συζητήσεις και τη βοήθειά της, όπως και για τη φιλοξενία της στο Μαγδεμβούργο. Με τους Ilaria Bartolini, Marco Patella και Καθ. Paolo Ciaccia, συνεργαστήκαμε στα πλαίσια του $PANDA$- Ήταν η πρώτη μου εξωτερική συνεργασία και έμαθα πολλά από αυτή, τους ευχαριστώ πολύ και για την εμπιστοσύνη που μου έδειξαν. Με τον Αλέξανδρο Καλούση συνεργαστήκαμε τυχαία, με αφετηρία μία πλατεία στο Πόρτο. Τον ευχαριστώ για τη συνεργασία και τη βοήθεια που μου προσέφερε στην κατανόηση των δέντρων απόφασης.

Θα ήθελα επίσης να ευχαριστήσω όλα τα μέλη του Εργαστηρίου Βάσεων Δεδομένων για τη συνεργασία και τη φιλία τους όλα αυτά τα χρόνια: τον Κώστα με τον οποίο μοιραστήκαμε αρκετούς προβληματισμούς, το Βαγγέλη με τον οποίο δουλέψαμε σε θέματα προτύπων, τους Μάκη, Νίκο, Ηλία με τους οποίους μπήκα στο χώρο των χωροχρονικών δεδομένων. Ευχαριστώ επίσης και τα νεότερα μέλη της ομάδας (Δέσποινα και Νίκο) που είναι πάντα πολύ ενθαρυντικοί.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους επιστήμονες με τους οποίους συνεργάστηκα ή απλώς συνομίλησα όλα αυτά τα χρόνια, καθώς επηρέασαν τη σκέψη μου και τις απόψεις μου σχετικά με το PhD και την ερευνητική ζωή γενικότερα. Αναφέρω ενδεικτικά το συμπόσιο στα Μαντάνεια, τις συναντήσεις στα πλαίσια του έργου PANDA, τα συνέδρια $HDMS$, το summer school στο Bolzano.

Επιπλέον, θα ήθελα να ευχαριστήσω τους φίλους μου που ήταν πάντα δίπλα μου και έδειξαν μεγάλη κατανόηση στα (πολλά) σκαμπανεβάσματα της διάθεσής μου. Πρόσθετα ευχαριστώ στο Σπύρο και για τη βοήθεια με το αγγλικό κείμενο. Και το Νίκο για τη βοήθεια με τη μετάφραση, και για όλα αυτά τα χρόνια που με στηρίζει και με βοηθάει ουσιαστικά.

Κλείνοντας, θα ήθελα να ευχαριστήσω τους γονείς μου και τις αδερφές μου που είναι πάντα δίπλα μου με την αγάπη και τη φροντίδα τους.

Ειρήνη Ντούτση

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Διαχείριση Προτύπων

Στο κεφάλαιο αυτό αναλύουμε την ανάγκη για διαχείριση προτύπων (*pattern management*) εστιάζοντας κυρίως σε μία συγκεκριμένη όψη του προβλήματος της διαχείρισης προτύπων, στο πρόβλημα της αποτίμησης της ομοιότητας μεταξύ προτύπων.

Το κεφάλαιο έχει οργανωθεί ως ακολούθως: Στην Ενότητα 1.1, εξηγούμε γιατί τα πρότυπα είναι τόσο δημοφιλή στις μέρες μας. Στην Ενότητα 1.2, παρουσιάζουμε περιληπτικά τα βασικά βήματα της Διαδικασίας Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (*Knowledge Discovery in Databases - KDD*) δίνοντας έμφαση στο βήμα της Εξόρυξης Γνώσης (*Data Mining - DM*). Στην Ενότητα 1.3, παρουσιάζουμε τα κίνητρα και τις ανάγκες που επιβάλλουν τη διαχείριση των προτύπων καθώς επίσης και μία επισκόπηση της ερευνητικής δουλειάς που έχει διεξαχθεί στον τομέα αυτό μέχρι τώρα. Στην Ενότητα 1.4, επικεντρωνόμαστε στο πρόβλημα της αποτίμησης της ομοιότητας μεταξύ προτύπων και παρουσιάζουμε τη σημαντικότητα αλλά και τις προκλήσεις που εγείρει η μελέτη του. Τέλος, στην Ενότητα 1.5, παρουσιάζουμε περιληπτικά τα θέματα που διαπραγματεύεται η συγκεκριμένη διδακτορική διατριβή.

Λέξεις κλειδιά διαχείριση προτύπων, Συστήματα Διαχείρισης Βάσεων Προτύπων (ΣΔΒΠ), ομοιότητα προτύπων.

1.1 Η επιτακτική ανάγκη για Εξόρυξη Γνώσης/ πρότυπα

Στις μέρες μας η ικανότητά μας να παράγουμε δεδομένα έχει αυξηθεί εντυπωσιακά εξαιτίας της ευρέως διαδεδομένης χρήσης των υπολογιστών σε κάθε τομέα της ζωής μας. Επιπλέον, λόγω της μεγάλης προόδου που έχει σημειωθεί στο συγκεκριμένο τομέα, έχουμε τη δυνατότητα να συλλέγουμε και να αποθηκεύουμε τα παραγόμενα δεδομένα. Ως αποτέλεσμα, *τεράστιες ποσότητες δεδομένων* συγκεντρώνονται από διάφορα πεδία εφαρμογών όπως οι επιχειρήσεις, οι επιστήμες, οι τηλεπικοινωνίες και ο κλάδος της υγείας. Επίσης, ο Παγκόσμιος Ιστός (*World Wide Web - WWW*) μας κατακλύζει με πληροφορίες. Σύμφωνα με πρόσφατη έρευνα [45] μάλιστα, ο κόσμος παράγει μεταξύ 1 και 2 exabytes μοναδικής πληροφορίας κάθε χρόνο, το οποίο αντιστοιχεί σε περίπου 250 megabytes για κάθε

άντρα, γυναίκα και παιδί στη γη.

Πέραν του τεράστιου τους όγκου, τα σύγχρονα δεδομένα χαρακτηρίζονται επίσης και από *χαμηλό επίπεδο λεπτομέρειας*. Για παράδειγμα, τα σουπερμάρκετ συλλέγουν δεδομένα σχετικά με τις συναλλαγές των πελατών τους, τα οποία περιλαμβάνουν τα προϊόντα που αγοράστηκαν από τους πελάτες σε κάθε συναλλαγή τους. Οι εταιρίες τηλεπικοινωνιών συλλέγουν δεδομένα από τα τηλεφωνήματα των χρηστών τους, τα οποία περιλαμβάνουν μεταξύ άλλων λεπτομέρειες όπως ο χρόνος και η διάρκεια μίας κλήσης. Οι ιδιοκτήτες δικτυακών τόπων συλλέγουν δεδομένα πλοήγησης από τους πελάτες τους, τα οποία περιλαμβάνουν λεπτομερείς πληροφορίες σχετικά με το χρόνο εισόδου σε μία σελίδα, τη διάρκεια παραμονής σε μία σελίδα κτλπ.. Οι συσκευές εντοπισμού θέσης, π.χ., GPS, μεταδίδουν κάθε λίγα λεπτά δεδομένα σχετικά με τη θέση του κινούμενου αντικειμένου και το χρόνο εντοπισμού του. Συνήθως τα δεδομένα αυτά αποκαλούνται *απλοϊκά δεδομένα (raw data)*, ακριβώς για να τονιστεί το χαμηλό επίπεδο λεπτομέρειας που αναπαριστούν.

Τα σύγχρονα δεδομένα χαρακτηρίζονται επίσης και από *μεγάλη ποικιλομορφία* (π.χ., συναλλαγές, εικόνες, μουσικά κομμάτια κτλπ.) και *μεγάλη πολυπλοκότητα* (π.χ., κείμενο, εικόνα, ήχος, βίντεο κτλπ.). Επιπλέον, τα δεδομένα στις μέρες μας παράγονται τόσο μέσω *συγκεντρωτικών (centralized)*, όσο και μέσω *κατανεμημένων (distributed)* τρόπων, γεγονός που επιβάλλει νέες προκλήσεις σχετικά με τη διαχείρισή τους.

Εξαιτίας των παραπάνω αναφερθέντων λόγων, είναι αδύνατο κάποιος άνθρωπος να εξερευνήσει εκτενώς και να εκμεταλλευτεί πλήρως αυτές τις τεράστιες συλλογές δεδομένων χωρίς τη χρήση κατάλληλων εργαλείων και μεθόδων. Η Διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (*Knowledge Discovery in Databases* - KDD) και η Εξόρυξη Γνώσης (*Data Mining -DM*) αποτελούν μία λύση στο συγκεκριμένο πρόβλημα καθώς εξάγουν από τα απλοϊκά δεδομένα γνώση, με τη μορφή προτύπων. Η Εξόρυξη Γνώσης αποτελεί στην πραγματικότητα ένα από τα βήματα της KDD διαδικασίας· αναφέρουμε περιληπτικά τα βήματα αυτά στην επόμενη ενότητα εστιάζοντας κυρίως στο DM βήμα.

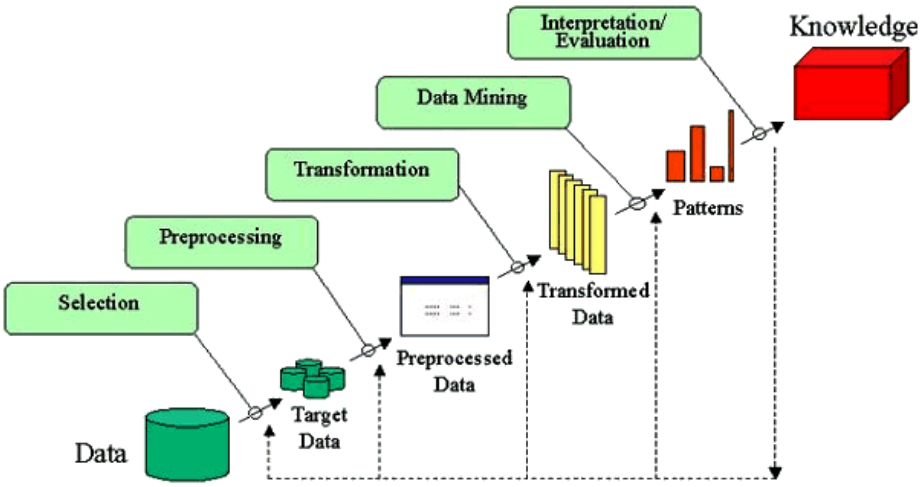
1.2 Η διαδικασία ΑΓΒΔ και η Εξόρυξη Γνώσης

Ορισμός 1 *Η διαδικασία Ανακάλυψης Γνώσης από Βάσεις Δεδομένων (ΑΓΒΔ) είναι μία μη-τετριμμένη διαδικασία για την εξαγωγή έγκυρων, προηγουμένως άγνωστων, ενδεχομένως χρήσιμων και ευκόλως κατανοητών προτύπων από τα δεδομένα [22].*

Ο παραπάνω ορισμός αποτελεί ένα γενικό ορισμό για τη διαδικασία ΑΓΒΔ. Ο όρος *δεδομένα* αντιστοιχεί σε ένα σύνολο στιγμιότυπων του υπό διερεύνηση προβλήματος. Θα μπορούσε να είναι, για παράδειγμα, ένα σύνολο από συναλλαγές πελατών στην περίπτωση ενός σουπερμάρκετ, ή ένα σύνολο εγγραφών-τηλεφωνημάτων στην περίπτωση μιας τηλεπικοινωνιακής εταιρίας. Ο όρος *πρότυπα* αναφέρεται σε εκφράσεις σε μία συγκεκριμένη γλώσσα που περιγράφουν κάποιο υποσύνολο των δεδομένων, π.χ., "*If income* $\leq 10K$*, then the loan request is rejected*". Τα πρότυπα που ανακαλύπτονται θα πρέπει να είναι *έγκυρα*, δηλαδή θα πρέπει να ισχύουν σε κάποιο βαθμό και για νέα, άγνωστα μέχρι στιγμής στιγμιότυπα του προβλήματος. Θα πρέπει επίσης να είναι πρωτότυπα, δηλαδή να αντικατοπτρίζουν κάποια νέα γνώση. Επίσης, τα πρότυπα θα πρέπει να είναι

χρήσιμα και να διευκολύνουν τον τελικό χρήστη στη λήψη αποφάσεων. Τέλος, αναμφίβολα θα πρέπει ο τελικός χρήστης να είναι σε θέση να ερμηνεύει τα πρότυπα κατά τρόπο *αποδοτικό* και *αποτελεσματικό*.

Ό όρος ΑΓΒΔ αναφέρεται σε όλη τη διαδικασία ανακάλυψης γνώσης από τα δεδομένα. Περιλαμβάνει πέντε βήματα: 1) *Επιλογή* των δεδομένων που σχετίζονται με το προς ανάλυση πρόβλημα, 2) *Προεπεξεργασία* των δεδομένων περιλαμβανόμενων εργασιών όπως ο καθαρισμός των δεδομένων και η ολοκλήρωσή τους, 3) *Μετατροπή* των δεδομένων σε μορφή κατάλληλη για εξόρυξη, 4) *Εξόρυξη Γνώσης* για την εξαγωγή προτύπων, και 5) *Ερμηνεία/Αξιολόγηση* των εξαγόμενων προτύπων προκειμένου να εντοπιστούν εκείνα τα πρότυπα που αναπαριστούν πραγματική γνώση με βάση κάποια μέτρα ενδιαφέροντος. Τα βήματα αυτά απεικονίζονται στο Σχήμα 1.1.



Σχήμα 1.1: Τα βήματα της διαδικασίας ΑΓΒΔ

Το βήμα της Εξόρυξης Γνώσης βρίσκεται στην καρδιά της διαδικασίας ΑΓΒΔ.

Ορισμός 2 *Η Εξόρυξη Γνώσης (ΕΓ) αποτελεί ένα βήμα της διαδικασίας ΑΓΒΔ και περιλαμβάνει την εφαρμογή αλγορίθμων ανάλυσης δεδομένων και Εξόρυξης Γνώσης που παράγουν μία συγκεκριμένη απαρίθμηση των προτύπων (ή μοντέλων) πάνω στα δεδομένα [22].*

Τα *πρότυπα* μπορούν να περιγραφούν ως συμπαγείς και πλούσιες σε σημασιολογία αναπαραστάσεις των πρωτογενών δεδομένων [70]: *συμπαγείς* καθώς παρέχουν μία υψηλού επιπέδου περιγραφή των χαρακτηριστικών των πρωτογενών δεδομένων και *πλούσια σε σημασιολογία* καθώς αποκαλύπτουν νέα γνώση που υπάρχει κρυμμένη στον όγκο των πρωτογενών δεδομένων.

Οι βασικοί στόχοι του βήματος της Εξόρυξης Γνώσης είναι η πρόβλεψη και η περιγραφή. Η *πρόβλεψη* αναφέρεται στην πρόβλεψη της τιμής ενός συγκεκριμένου γνωρίσματος και εφαρμόζεται πάνω σε νέα, άγνωστα μέχρι στιγμής στιγμιότυπα του προβλήματος, ενώ η *περιγραφή* αναφέρεται στην εξαγωγή εύκολα ερμηνεύσιμων προτύπων από τα δεδομένα.

Οι βασικές εργασίες της Εξόρυξης Γνώσης είναι: η ταξινόμηση, η συσταδοποίηση και η εξαγωγή κανόνων συσχέτισης. Η *ταξινόμηση (classification)* αποσκοπεί στη δημιουργία ενός μοντέλου που διαχωρίζει τα δεδομένα σε προκαθορισμένες κλάσεις. Το μοντέλο αυτό χρησιμοποιείται εν συνεχεία για την πρόβλεψη της κλάσης αντικειμένων για τα οποία η κλάση τους δεν είναι γνωστή. Ένα κλασσικό παράδειγμα ταξινόμησης είναι αυτό της έγκρισης δανείων σε μία τράπεζα, όπου ο στόχος είναι να προβλεφθεί αν μία αίτηση για δάνειο πρέπει να εγκριθεί ή όχι. Η *συσταδοποίηση(clustering)* αναφέρεται στον διαμερισμό των δεδομένων σε ομάδες/ συστάδες έτσι ώστε παρόμοια αντικείμενα να τοποθετούνται στην ίδια ομάδα. Ένα παράδειγμα συσταδοποίησης, είναι ο διαμερισμός των ανθρώπων σε ομάδες με βάση τη μόρφωσή τους και το βιοτικό τους επίπεδο. Η *εξαγωγή κανόνων συσχέτισης (association rules extraction)* αποσκοπεί στην ανακάλυψη συσχετίσεων μεταξύ ζευγών της μορφής γνώρισμα-τιμή που εμφανίζονται συχνά μαζί στα δεδομένα. Ένα πολύ γνωστό πεδίο εφαρμογής είναι τα σουπερμάρκετ όπου ο στόχος είναι η ανακάλυψη των προϊόντων που οι καταναλωτές τείνουν να αγοράζουν μαζί.

Ανάμεσα στα πιο δημοφιλή μοντέλα Εξόρυξης Γνώσης (ή τύπους προτύπων) ανήκουν τα δέντρα απόφασης, οι κανόνες συσχέτισης, οι συστάδες και οι ακολουθίες.

1.3 Διαχείριση προτύπων

Εξαιτίας της ευρείας εφαρμογής της διαδικασίας ΑΓΒΔ και ως αποτέλεσμα της πλημμύρας δεδομένων που παρατηρείται στις μέρες μας, ο όγκος των προτύπων που εξάγονται σήμερα από ετερογενείς πηγές είναι *τεράστιος* και, συχνά, *μη-διαχειρίσιμος* από τους ανθρώπους. Συνεπώς, υπάρχει ανάγκη για διαχείριση προτύπων που περιλαμβάνει θέματα μοντελοποίησης, αποθήκευσης, ανάκτησης και επερωτήσεων. Η διαχείριση προτύπων δεν είναι εύκολη καθώς εκτός από τις μεγάλες ποσότητες των παραγόμενων προτύπων, υπάρχει και μεγάλη ποικιλία όσον αφορά στους τύπους των προτύπων, γεγονός που οφείλεται στις διαφορετικές ανάγκες εφαρμογών που κάθε τύπος προσπαθεί να καλύψει.

Μέχρι τώρα, η πλειοψηφία των εργασιών στον τομέα της Εξόρυξης Γνώσης εστιάζει κυρίως σε αλγορίθμους και τεχνικές για την αποδοτική εξαγωγή των προτύπων, αδιαφορώντας κατά κάποιο τρόπο για το θέμα της διαχείρισης των εξαγόμενων προτύπων. Τελευταία, ωστόσο, η ανάγκη για διαχείριση προτύπων έχει αναγνωριστεί τόσο από την επιστημονική κοινότητα όσο και από τον τομέα της βιομηχανίας/ αγοράς με αποτέλεσμα να έχουν προταθεί κάποιες προσεγγίσεις για την αποδοτική διαχείριση των προτύπων. Η βασική διαφορά ανάμεσα στις επιστημονικές και τις βιομηχανικές προσεγγίσεις είναι ότι οι επιστημονικές προσεγγίσεις προσπαθούν να δώσουν συνολικές λύσεις στο πρόβλημα της διαχείρισης αντιμετωπίζοντας όλα τα επιμέρους θέματα, ενώ οι βιομηχανικές προσεγγίσεις εστιάζουν κυρίως σε θέματα αναπαράστασης και αποθήκευσης προτύπων αποσκοπώντας στην εύκολη ανταλλαγή προτύπων μεταξύ διαφορετικών εφαρμογών.

Στη συνέχεια, παρουσιάζουμε περιληπτικά αυτές τις προσεγγίσεις. Για περαιτέρω λεπτομέρειες βλέπε [61, 76].

1.3.1 Επιστημονικές προσεγγίσεις

Οι επιστημονικές προσεγγίσεις προσπαθούν να δώσουν ολοκληρωμένες λύσεις στο πρόβλημα της διαχείρισης προτύπων παρέχοντας δυνατότητες αναπαράστασης, αποθήκευσης και ανάκτησης προτύπων. Οι βασικές προσεγγίσεις σε αυτή την κατηγορία είναι οι επαγωγικές βάσεις δεδομένων, το μοντέλο των 3–Κόσμων και το PBMS μοντέλο.

Επαγωγικές Βάσεις Δεδομένων Οι επαγωγικές βάσεις δεδομένων (inductive databases)[33], πρωτοπαρουσιάστηκαν το 1996 και στηρίζονται στην ιδέα ότι η διαδικασία Εξόρυξης Γνώσης θα πρέπει να υποστηρίζεται από την τεχνολογία βάσεων δεδομένων. Αυτός είναι ο λόγος που οι επαγωγικές βάσεις δεδομένων βασίζονται σε μία ολοκληρωμένη αρχιτεκτονική όπου τα δεδομένα και τα πρότυπα αποθηκεύονται στην ίδια αποθήκη.

Στις επαγωγικές βάσεις δεδομένων, η διαδικασία ΑΓΒΔ θεωρείται ως εκτεταμένη επεξεργασία επερωτήσεων στην οποία οι χρήστες μπορούν να θέτουν ερωτήσεις τόσο στα δεδομένα όσο και στα πρότυπα. Για το σκοπό αυτό, μία αποκαλούμενη *επαγωγική γλώσσα επερωτήσεων* χρησιμοποιείται, η οποία αποτελεί επέκταση μιας γλώσσας επερωτήσεων βάσεων δεδομένων που επιτρέπει σε κάποιον ι) να επιλέγει, να διαχειρίζεται και να ρωτά *δεδομένα* όπως στις κλασσικές επερωτήσεις ιι) να επιλέγει, να διαχειρίζεται και να ρωτά *πρότυπα* και, ιιι) να εκτελεί *cross–over* ερωτήματα πάνω στα πρότυπα, δηλαδή, ερωτήματα που συσχετίζουν τα πρότυπα με τα πρωτογενή δεδομένα. Λόγω της σημαντικότητας των επερωτήσεων στις επαγωγικές βάσεις δεδομένων, διάφορες γλώσσες επερωτήσεων-επεκτάσεις της SQL έχουν προταθεί, όπως η DMQL [29], η MINE–RULE [48] και η MINE–SQL [34].

Το μοντέλο των 3-Κόσμων Το μοντέλο των 3-Κόσμων (3-Worlds) [38], το οποίο προτάθηκε το 2000, παρέχει ένα ενοποιημένο πλαίσιο για τη διαχείριση προτύπων και αποτελεί την πρώτη προσπάθεια διαχωρισμού της διαχείρισης των προτύπων από τη διαχείριση των πρωτογενών δεδομένων.

Το μοντέλο των 3-Κόσμων στηρίζεται σε μία ξεχωριστή αρχιτεκτονική που αποτελείται από τρεις διακριτούς κόσμους: 1) τον *κόσμο των προτύπων* (intensional world) 2) τον *κόσμο των δεδομένων* (data world) και, 3) τον ενδιάμεσο κόσμο (extensional world) που καθορίζει την αντιστοίχιση μεταξύ των προτύπων και των δεδομένων.

Ο χειρισμός των επιμέρους κόσμων μπορεί να γίνει με κάποια άλγεβρα επιλογής. Για τον κόσμο των προτύπων, οι συγγραφείς προτείνουν την *άλγεβρα διαστάσεων* (dimension algebra) που επεκτείνει παραδοσιακούς τελεστές της σχεσιακής άλγεβρας μέσω νέων τελεστών που προσδίδουν μεγάλη αξία στην Εξόρυξη Γνώσης και στην ανάλυση των δεδομένων. Για να διευκολύνουν την κίνηση μέσα και έξω από τους κόσμους, οι συγγραφείς προτείνουν τους λεγόμενους *τελεστές γεφύρωσης* (bridge operators), και συγκεκριμένα τους τελεστές populate, mine, lookup και refresh.

Το PBMS μοντέλο Το PBMS (Pattern Base Management Systems) μοντέλο, το οποίο προτάθηκε το 2003 στα πλαίσια του Ευρωπαϊκού έργου PANDA [65], παρέχει ένα ενοποιημένο πλαίσιο για την αναπαράσταση προτύπων και στηρίζεται σε μία ξεχωριστή αρχιτεκτονική όπου τα δεδομένα και τα πρότυπα αποθηκεύονται σε διαφορετικές αποθήκες.

Οι διαφορετικοί τύποι προτύπων αναπαρίστανται κάτω από ένα ενιαίο σχήμα με τη μορφή $pt = (n, ss, ds, ms, et)$: n είναι το *όνομα* του τύπου προτύπων· ss είναι το *σχήμα της δομής* (structure schema) που ορίζει το χώρο των προτύπων· ds είναι το *σχήμα της προέλευσης* (source schema) που ορίζει το χώρο των δεδομένων προέλευσης· ms είναι το *σχήμα των μέτρων* (measure schema) που περιγράφει τα μέτρα που ποσοτικοποιούν την ποιότητα της αναπαράστασης των πρωτογενών δεδομένων που επιτυγχάνουν τα πρότυπα· et είναι το *πρότυπο έκφρασης* (expression template) που περιγράφει τη σχέση μεταξύ του χώρου προέλευσης και του χώρου των προτύπων, μεταφέροντας τη σημασιολογία των προτύπων. Ένα *πρότυπο* (pattern) αποτελεί ένα στιγμιότυπο ενός τύπου προτύπων. Πρότυπα με παρόμοια σημασιολογία ομαδοποιούνται σε *κλάσεις* (classes).

Για το χειρισμό των προτύπων, μία Γλώσσα Χειρισμού Προτύπων (Pattern Manipulation Language - PML) και μία Γλώσσα Επερωτήσεων Προτύπων (Pattern Query Language - PQL) έχουν προταθεί που υποστηρίζουν επερωτήσεις πάνω στα πρότυπα και cross over επερωτήσεις μεταξύ προτύπων και πρωτογενών δεδομένων. Το βασικό μοντέλο [70] έχει επεκταθεί προσθέτοντας στον ορισμό των προτύπων την έννοια της προσωρινής εγκυρότητας (temporal validity), της σημασιολογικής εγκυρότητας (semantic validity) και της ασφάλειας (safety) [15].

1.3.2 Βιομηχανικές προσεγγίσεις

Οι βιομηχανικές προσεγγίσεις εστιάζουν κυρίως σε θέματα αναπαράστασης και αποθήκευσης προτύπων στοχεύοντας κυρίως στην εύκολη ανταλλαγή προτύπων μεταξύ διαφορετικών εφαρμογών παρά στην αποδοτική διαχείριση των εξαγόμενων προτύπων. Διάφορες προδιαγραφές και πρότυπα (PMML, CWM, ISO SQL/MM, JDMAPI) έχουν προταθεί σε αυτή την κατηγορία, μερικά από τα οποία παρουσιάζονται πιο αναλυτικά στη συνέχεια. Επίσης, υπάρχουν επεκτάσεις υπαρχόντων εμπορικών συστημάτων όπως το Oracle Data Mining [63], IBM DB2 Intelligent Miner [32], Microsoft SQL Server Analysis Manager [52].

Predictive Model Markup Language (PMML) Η PMML [20] είναι μία γλώσσα βασιζόμενη στην XML που παρέχει στις εταιρίες ένα γρήγορο και εύκολο τρόπο για να ορίσουν μοντέλα Εξόρυξης Γνώσης και στατιστικά μοντέλα χρησιμοποιώντας μία μέθοδο ανεξάρτητη του κατασκευαστή και να μοιράζονται αυτά τα μοντέλα μεταξύ PMML συμβατών εφαρμογών. Η δομή των μοντέλων περιγράφεται μέσω ενός XML σχήματος.

ISO SQL/MM Η SQL/MM [77] αφορά προδιαγραφές για την υποστήριξη της διαχείρισης κοινών τύπων δεδομένων (κείμενα, εικόνες, αποτελέσματα Εξόρυξης Γνώσης, ...) που σχετίζονται με πολυμεσικές εφαρμογές. Το τμήμα 6 της SQL/MM αναφέρεται στην Εξόρυξη Γνώσης. Η SQL/MM ορίζει πρώτης κλάσης SQL τύπους που μπορούν να προσπελαστούν μέσω της βασικής σύνταξης της SQL:1999. Στην περίπτωση των μοντέλων Εξόρυξης Γνώσης, κάθε μοντέλο έχει ένα αντίστοιχο τύπο ορισμένο από το χρήστη και δομημένο με βάση την SQL.

Common Warehouse Model (CWM) Το CWM [17] περιλαμβάνει προδιαγραφές που επιτρέπουν την εύκολη ανταλλαγή μεταδεδομένων μεταξύ εργαλείων για αποθήκες δεδομένων και αποθηκών από μεταδεδομένα σε κατανεμημένα

ετερογενή περιβάλλοντα. Αποτελείται από ένα πλήθος υπο-μεταμοντέλων που αναπαριστούν κοινά μεταδεδομένα αποθηκών στις περιοχές της ανάλυσης δεδομένων (OLAP, Εξόρυξη Γνώσης, ...) και της διαχείρισης αποθηκών δεδομένων (Warehouse Management).

Java Data Mining API (JDMAPI) Το Java Data Mining API [37] αντιμετωπίζει την ανάγκη για ένα Java API ανεξάρτητο από το εκάστοτε σύστημα, το οποίο θα υποστηρίζει τη δημιουργία, την αποθήκευση, την προσπέλαση και την συντήρηση των δεδομένων και των μεταδεδομένων. Παρέχει μία τυποποιημένη πρόσβαση στα πρότυπα της Εξόρυξης Γνώσης που αναπαρίστανται σε διάφορα σχήματα, π.χ., PMML [20].

1.4 Ομοιότητα προτύπων

Μέχρι στιγμής αναφερθήκαμε στην σημαντικότητα της διαχείρισης προτύπων, τα διαφορετικά θέματα που η έννοια της διαχείρισης περιλαμβάνει (π.χ., μοντελοποίηση, αποθήκευση και χειρισμός των προτύπων) και την πρόοδο που έχει συντελεστεί μέχρι στιγμής προς αυτή την κατεύθυνση τόσο από την επιστημονική κοινότητα όσο και από την πλευρά της βιομηχανίας.

Ανάμεσα στις διάφορες ενδιαφέρουσες λειτουργίες που μπορούν να οριστούν πάνω στα πρότυπα, μία από τις πιο σημαντικές λειτουργίες είναι αυτή της αποτίμησης της ομοιότητας μεταξύ προτύπων. Στη συνέχεια, παρουσιάζουμε διάφορες εφαρμογές που δηλώνουν τη σημαντικότητα του προβλήματος της αποτίμησης της ομοιότητας μεταξύ προτύπων και τις προκλήσεις που εγείρει η αντιμετώπιση του εν λόγω προβλήματος.

1.4.1 Η σημαντικότητα του προβλήματος της ομοιότητας προτύπων

Ένα πλήθος ενδιαφερουσών εφαρμογών προκύπτει από τον ορισμό της ομοιότητας/απόστασης μεταξύ των προτύπων. Στη συνέχεια, περιγράφουμε εν συντομία μερικές από τις εφαρμογές αυτές :

- *Ερωτήματα Ομοιότητας* Η άμεση εφαρμογή ενός τελεστή ομοιότητας είναι στον ορισμό επερωτήσεων ομοιότητας πάνω σε ένα σύνολο από πρότυπα (ή βάσεις προτύπων), συμπεριλαμβανομένων των επερωτήσεων k-πλησιέστερων γειτόνων (k-nearest neighbor queries) (δηλαδή, βρες τα k πιο όμοια πρότυπα σε σχέση με ένα δοθέν πρότυπο) και των επερωτήσεων εύρους τιμών (range queries) (δηλαδή, βρες τα ποιο όμοια πρότυπα σε σχέση με ένα δοθέν πρότυπο τα οποία βρίσκονται σε συγκεκριμένα όρια τιμών όσον αφορά στα γνωρίσματά τους). Ο αποδοτικός υπολογισμός της ομοιότητας είναι ένα από τα βασικά θέματα σε ένα ΣΔΒΠ [70] με εφαρμογές σε θέματα ευρετηριοποίησης και ανάκτησης προτύπων.

- *Παρακολούθηση και εντοπισμός αλλαγών* Μία άλλη εφαρμογή της ομοιότητας είναι η παρακολούθηση και ο εντοπισμός αλλαγών στα πρότυπα, π.χ., βρες τις αλλαγές στη συμπεριφορά των πελατών στην πάροδο του χρόνου. Αυτό είναι εξαιρετικά σημαντικό στις μέρες μας που τα δεδομένα είναι δυναμικά και συνεπώς, τα εξαγόμενα πρότυπα είναι επίσης δυναμικά, και εμπεριέχουν

την έννοια του χρόνου. Επιπλέον, είναι συνήθως πιο βοηθητικό για τον τελικό χρήστη να γνωρίζει πως τα παλιά πρότυπα (που αντιπροσωπεύουν την τρέχουσα γνώση π.χ., της εταιρίας για το πρόβλημα) έχουν εξελιχθεί, παρά να έχει να διαχειριστεί ένα ακόμη σύνολο προτύπων (το οποίο αναπαριστά τη νέα γνώση π.χ., της εταιρίας σχετικά με το πρόβλημα).

Ο εντοπισμός αλλαγών είναι επίσης χρήσιμος για τον συγχρονισμό των προτύπων στη διαδικασία ΑΓΒΔ έτσι ώστε να αντανακλούν πάντα τα πρωτογενή δεδομένα από τα οποία έχουν εξαχθεί (π.χ., ενημέρωσε τα πρότυπα μόνο όταν τα αντίστοιχα πρωτογενή δεδομένα έχουν αλλάξει σημαντικά).

- *Σύγκριση συνόλων δεδομένων* Μία κοινή τεχνική για τη σύγκριση δύο συνόλων δεδομένων είναι μέσω της σύγκρισης των συνόλων προτύπων που εξάγονται από τα δεδομένα αυτά, π.χ., σύγκρινε δύο σουπερμάρκετ με βάση την αγοραστική συμπεριφορά των πελατών τους (η οποία μπορεί να αποδοθεί μέσω κάποιου μοντέλου συσταδοποίησης).

  Η παραπάνω συσχέτιση γίνεται με τη λογική ότι καθώς τα πρότυπα διατηρούν σε κάποιο βαθμό μέρος της πληροφορίας που υπάρχει στα πρωτογενή δεδομένα, θα μπορούσε η ομοιότητα των προτύπων να χρησιμοποιηθεί ως μέτρο της ομοιότητας των πρωτογενών δεδομένων.

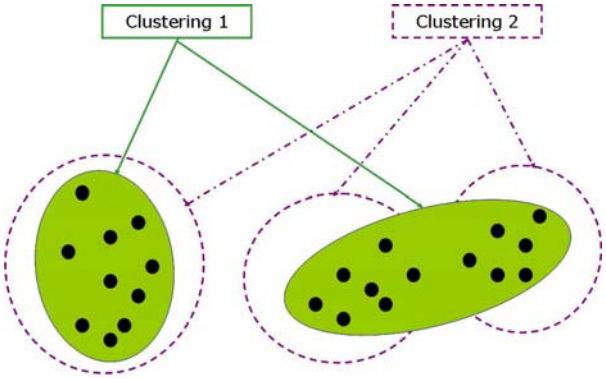
  Η εύρεση μίας τέτοιας (ακριβούς ή προσεγγιστικής) αντιστοίχισης μεταξύ της ομοιότητας των προτύπων και της ομοιότητας των πρωτογενών δεδομένων είναι εξαιρετικά χρήσιμη. Με τον τρόπο αυτό, θα μπορούσαμε, για παράδειγμα να αποφύγουμε το δύσκολο έργο της σύγκρισης συνόλων δεδομένων όταν υπάρχουν διαθέσιμα τα αντίστοιχα σύνολα προτύπων ή ακόμα και να αποφύγουμε να εξάγουμε πρότυπα από ένα σύνολο δεδομένων όταν αυτό είναι παρόμοιο με κάποιο άλλο σύνολο δεδομένων για το οποίο έχουμε ήδη εξάγει τα πρότυπα.

- *Αξιολόγηση αλγορίθμων Εξόρυξης Γνώσης*

  Συνήθως η σύγκριση μεταξύ δύο αλγορίθμων (ή διαφορετικών παραμέτρων του ίδιου αλγορίθμου) περιορίζεται στην οπτική αντιπαράθεση των αποτελεσμάτων τους. Ωστόσο, μία τέτοια αξιολόγηση θα μπορούσε να γίνει αυτόματα συγκρίνοντας τις εξόδους/αποτελέσματά τους (Σχήμα 1.2). Ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων (π.χ., πυκνό - αραιό) ένας αλγόριθμος μπορεί να είναι πιο κατάλληλος σε σχέση με κάποιον άλλο. Συνεπώς, κάποιος μπορεί να χρειαστεί να πειραματιστεί με πολλούς αλγορίθμους και να αξιολογήσει τα αποτελέσματά τους προκειμένου να καταλήξει στον αλγόριθμο που είναι πιο κατάλληλος για τις ανάγκες του.

- *Εξόρυξη Γνώσης που διατηρεί την ιδιωτικότητα των δεδομένων* Υπάρχουν περιπτώσεις όπου εξαιτίας της ιδιωτικότητας (privacy) των δεδομένων μόνο τα πρότυπα είναι διαθέσιμα και όχι τα πρωτογενή δεδομένα. Ανάλογα με τις παραμέτρους της Εξόρυξης Γνώσης, είναι δύσκολο να επανακτήσει κανείς τα αρχικά δεδομένα (για παράδειγμα, η απόφαση αν ένα σύνολο δεδομένων είναι συμβατό με ένα σύνολο από συχνά στοιχειοσύνολα είναι $NP-hard$ [49]). Στην περίπτωση αυτή, θα πρέπει να στηριχθούμε στα διαθέσιμα σύνολα προτύπων προκειμένου να συγκρίνουμε τα σύνολα δεδομένων.

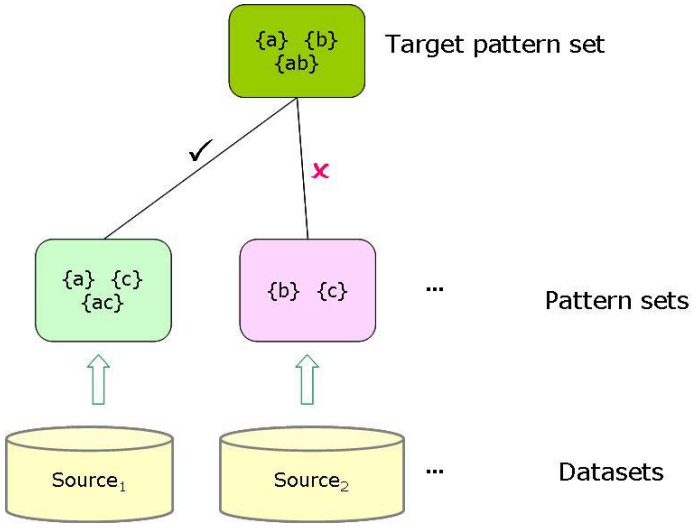
- *Εξόρυξη Γνώσης από κατανεμημένες πηγές δεδομένων* Σε ένα κατανεμημένο περιβάλλον, η Εξόρυξη Γνώσης δεν είναι συγκεντρωτική (centralized) καθώς



Σχήμα 1.2: Σύγκριση των αποτελεσμάτων διαφορετικών αλγορίθμων συσταδοποίησης πάνω στο ίδιο σύνολο δεδομένων

πρότυπα που είναι ισχυρά σε τοπικό επίπεδο μπορεί να χαθούν. Μία λύση θα μπορούσε να είναι η ομαδοποίηση των συνόλων δεδομένων σε ομάδες παρόμοιων δεδομένων, και εν συνεχεία η Εξόρυξη Γνώσης πάνω σε κάθε ομάδα ξεχωριστά [44].

- *Ανακάλυψη ακραίων τιμών ή μη αναμενόμενων προτύπων* Η ανακάλυψη ακραίων τιμών (outlier) ή μη αναμενόμενων προτύπων αποτελεί μία ακόμη εφαρμογή της ομοιότητας με ιδιαίτερο ενδιαφέρον για τον τελικό χρήστη. Μία πιθανή λύση στο πρόβλημα είναι να συγκρίνουμε τα εξαγόμενα πρότυπα με κάποιο πρότυπο-στόχο, το οποίο θα μπορούσε να καθορίζει ο χρήστης (Σχήμα 1.3). Τα πρότυπα που διαφέρουν σημαντικά από το πρότυπο-στόχο μπορούν να θεωρηθούν ως ακραία.



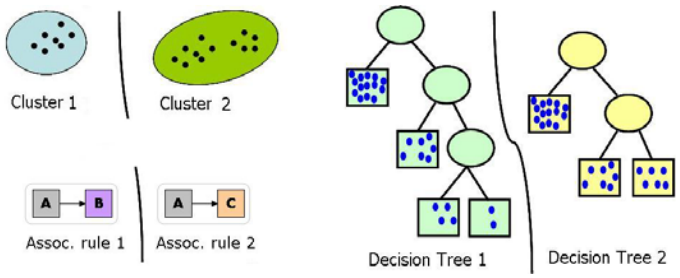
Σχήμα 1.3: Σύγκριση των εξαγόμενων προτύπων σε σχέση με ένα πρότυπο-στόχο

1.4.2 Οι προκλήσεις του προβλήματος της ομοιότητας προτύπων

Στις προηγούμενες ενότητες, παρουσιάσαμε ένα σύνολο από παραδείγματα που δείχνουν την σημαντικότητα του προβλήματος της αποτίμησης της ομοιότητας μεταξύ προτύπων και παρακινούν την περαιτέρω διερεύνηση του εν λόγω προβλήματος.

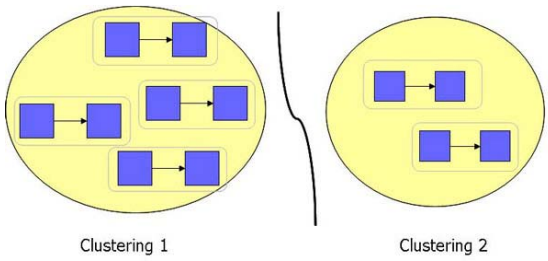
Ωστόσο, η αποτίμηση της ομοιότητας των προτύπων δεν είναι ένα εύκολο έργο όπως θα φανεί και από τις προκλήσεις που εγείρει, τις οποίες και παρουσιάζουμε στη συνέχεια:

Πρόκληση 1 *Πρώτα από όλα, θα πρέπει να οριστούν* τελεστές ομοιότητας για τους διάφορους τύπους προτύπων, *π.χ., συσταδοποιήσεις (Σχήμα 1.4, αριστερά), δέντρα απόφασης (Σχήμα 1.4, δεξιά), κανόνες συσχέτισης, συχνά στοιχειοσύνολα κτλπ.*



Σχήμα 1.4: Συγκρίνοντας συσταδοποιήσεις (αριστερά), δέντρα απόφασης (δεξιά)

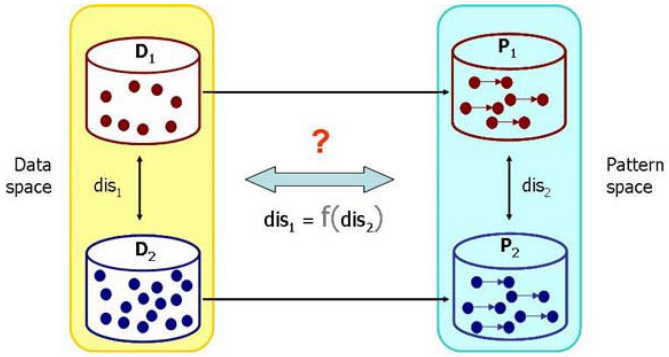
Πρόκληση 2 *Εκτός από τα πρότυπα που ορίζονται πάνω σε πρωτογενή δεδομένα (αποκαλούμενα και απλά πρότυπα), υπάρχουν επίσης και πρότυπα που ορίζονται πάνω σε άλλα πρότυπα, π.χ., μία συστάδα από κανόνες συσχέτισης (Σχήμα 1.5), ένας κανόνας συσχέτισης από συστάδες κτλπ.. Για τα πρότυπα αυτά (αποκαλούμενα σύνθετα πρότυπα, θα πρέπει επίσης να οριστούν τελεστές ομοιότητας.*



Σχήμα 1.5: Σύγκριση συστάδων από κανόνες συσχέτισης

Πρόκληση 3 *Μία ακόμα ενδιαφέρουσα άποψη του προβλήματος της αποτίμησης της ομοιότητας είναι η σημασιολογία των τελεστών ανομοιότητας σε σχέση με τα*

πρωτογενή δεδομένα από τα οποία έγινε η εξαγωγή των προτύπων. Πιο συγκεκριμένα, είναι πραγματικά ενδιαφέρον να εξερευνήσει κανείς κατά πόσο η ομοιότητα στο χώρο των προτύπων σχετίζεται με την ομοιότητα στο χώρο των πρωτογενών δεδομένων, από τον οποίο προέρχονται τα πρότυπα (Σχήμα 1.6).



Σχήμα 1.6: Συσχετισμός της ομοιότητας στο χώρο των προτύπων με την ομοιότητα στο χώρο των πρωτογενών δεδομένων

Στα πλαίσια αυτής της διατριβής, αντιμετωπίζουμε τις παραπάνω προκλήσεις για κάποιους από τους πιο δημοφιλείς τύπους προτύπων, συγκεκριμένα για συχνά στοιχειοσύνολα, δέντρα απόφασης και συστάδες, ενώ προτείνουμε επίσης και μεθόδους και τεχνικές για την αποτίμηση της ομοιότητας μεταξύ προτύπων αυθαίρετης πολυπλοκότητας όπως οι δικτυακοί τόποι και οι γράφοι.

1.5 Περιεχόμενα διατριβής

Η σημασία του προβλήματος της διαχείρισης των προτύπων και ειδικότερα του προβλήματος της αποτίμησης της ομοιότητας μεταξύ προτύπων προκύπτει ξεκάθαρα από το κεφάλαιο αυτό. Αν και πρόσφατα, αρκετές εργασίες έχουν παρουσιαστεί στον τομέα της διαχείρισης προτύπων, οι εργασίες αυτές εστιάζουν κατά κύριο λόγο στην αναπαράσταση, αποθήκευση και ανάκτηση των προτύπων. Πολλά ενδιαφέροντα ζητήματα παραμένουν προς μελέτη, όπως για παράδειγμα η αποτίμηση της ομοιότητας μεταξύ προτύπων, θέματα ευρετηριοποίησης, θέματα οπτικοποίησης κτλπ.. Στην παρούσα διατριβή, ερευνάμε τις διαφορετικές πτυχές του προβλήματος της ανομοιότητας μεταξύ προτύπων για κάποιους από τους πιο δημοφιλής τύπους προτύπων, συγκεκριμένα για συχνά στοιχειοσύνολα (frequent itemsets), δέντρα απόφασης (decision trees) και συστάδες (clusters). Πρώιμες εκδόσεις της συζήτησης που παρουσιάζεται σε αυτό το κεφάλαιο εμφανίζονται στις [58, 59, 40, 76, 61].

Στο Κεφάλαιο 2, κάνουμε μια εισαγωγή στους βασικούς τύπους προτύπων που μελετώνται μέσω αυτής της διατριβής (συχνά στοιχεοσύνολα, δέντρα απόφασης και συστάδες), ώστε ο αναγνώστης να είναι εξοικειωμένος με αυτές τις έννοιες. Στο ίδιο κεφάλαιο, παρουσιάζουμε επίσης ένα σχήμα αναπαράστασης για τα πρότυπα, το οποίο στηρίζεται τόσο στα δεδομένα από τα οποία έγινε η εξαγωγή των προτύπων (*extensionaldescription*) όσο και στην έννοια/ νόημα που αναπαριστούν

τα πρότυπα (*intensionaldescription*). Μέρος αυτής της δουλειάς εμφανίζεται στις [56, 60].

Στο Κεφάλαιο 3, παρουσιάζουμε το πλαίσιο $PANDA$ για την αποτίμηση της ανομοιότητας μεταξύ προτύπων αυθαίρετης πολυπλοκότητας. Το $PANDA$ είναι ικανό να διαχειριστεί τόσο πρότυπα που ορίζονται πάνω σε πρωτογενή δεδομένα (απλά πρότυπα) όσο και πρότυπα που ορίζονται πάνω σε άλλα πρότυπα (σύνθετα πρότυπα). Εφαρμόζουμε το $PANDA$ για βασικούς τύπους προτύπων, συμπεριλαμβανομένων των συχνών στοιχειοσυνόλων, των δέντρων απόφασης και των συστάδων, και για πιο σύνθετους τύπους προτύπων, όπως οι συλλογές κειμένων και οι ιστοσελίδες. Αυτό το μέρος είναι συμβατό με την Πρόκληση 2. Μία πρώιμη έκδοση αυτής της μελέτης έχει δημοσιευτεί στην [8], ενώ μία εκτεταμένη έκδοση έχει υποβληθεί στην [9].

Στο Κεφάλαιο 4, επικεντρωνόμαστε στο πρόβλημα της αποτίμησης της ομοιότητας μεταξύ συνόλων από συχνά στοιχειοσύνολα. Μελετάμε πώς οι διαφορετικές παράμετροι της Εξόρυξης Γνώσης, και συγκεκριμένα το κατώφλι $minSupport$ που χρησιμοποιείται για την παραγωγή των συχνών στοιχειοσυνόλων και η αναπαράσταση του πλέγματος των στοιχειοσυνόλων (itemsets lattice) (συγκεκριμένα, συχνά στοιχειοσύνολα (frequent itemsets), κλειστά συχνά στοιχειοσύνολα (closed frequent itemsets) ή μέγιστα συχνά στοιχειοσύνολα (maximal frequent itemsets», επηρεάζουν το παραγόμενο αποτέλεσμα ομοιότητας μεταξύ δύο συνόλων από στοιχειοσύνολα. Αυτό το κεφάλαιο είναι συμβατό με τις Προκλήσεις1 και 3. Μία πρώιμη έκδοση αυτής της μελέτης έχει δημοσιευτεί στις [54, 56], ενώ μία εκτεταμένη έκδοση έχει υποβληθεί στην [62].

Στο Κεφάλαιο 5, επικεντρωνόμαστε στην αξιολόγηση της ομοιότητας μεταξύ δέντρων απόφασης και συνόλων δεδομένων κατηγοριοποίησης. Συγκεκριμένα, παρουσιάζουμε ένα γενικό πλαίσιο για την αποτίμηση της ομοιότητας το οποίο βασίζεται στα δέντρα απόφασης και περιλαμβάνει ως επιμέρους περιπτώσεις την αποτίμηση της σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης και την αποτίμηση της ομοιότητας μεταξύ συνόλων δεδομένων κατηγοριοποίησης με βάση διαφορετικές κατανομές πυκνότητας πιθανότητας πάνω στο χώρο του προβλήματος. Αυτό το κεφάλαιο είναι συμβατό με τις Προκλήσεις 1 και 3. Τα αποτελέσματα αυτής της μελέτης έχουν δημοσιευτεί στις [55, 56].

Στο Κεφάλαιο 6, επικεντρωνόμαστε στην παρακολούθηση της εξέλιξης ενός δυναμικού πληθυσμού στην πορεία του χρόνου με τη βοήθεια των μοντέλων συσταδοποίησης που εξάγονται από αυτόν τον πληθυσμό. Πιο συγκεκριμένα, προτείνουμε τις διαφορετικές μεταβολές που μια συστάδα μπορεί να υποστεί και προτείνουμε μεθόδους για τον εντοπισμό αυτών των μεταβολών. Επιπλέον, μελετάμε πώς η εξέλιξη του πληθυσμού μπορεί να οργανωθεί αποτελεσματικά προκειμένου ο τελικός χρήστης να αποκτήσει καλύτερη κατανόηση των δεδομένων του και της μεταβολής τους στο χρόνο. Αυτό το κεφάλαιο είναι συμβατό με τις Προκλήσεις 1 και 3. Διάφορα τμήματα αυτής της μελέτης έχουν δημοσιευτεί στις [74, 73, 75, 72, 56], ενώ μία εκτεταμένη έκδοση έχει υποβληθεί στην [57].

Ολοκληρώνουμε την παρούσα διατριβή στο Κεφάλαιο 7, όπου συζητάμε επίσης τα ανοικτά ζητήματα και πιθανές επεκτάσεις.

Κεφάλαιο 2

Η Έννοια των Προτύπων στην Εξόρυξη Γνώσης

Στο κεφάλαιο αυτό παρουσιάζουμε τρεις δημοφιλείς τύπους προτύπων Εξόρυξης Γνώσης, τα συχνά στοιχειοσύνολα (frequent itemsets) και τους κανόνες συσχέτισης (association rules), τις συστάδες (clusters) και τις συσταδοποιήσεις (clusterings) και τα δέντρα απόφασης (decision trees). Στόχος του κεφαλαίου είναι η εξοικείωση του τελικού αναγνώστη με την έννοια των προτύπων στην Εξόρυξη Γνώσης.

Πριν όμως την παρουσίαση των βασικών τύπων προτύπων περιγράφουμε (Ενότητα 2.2) ένα σχήμα αναπαράστασης για τα πρότυπα το οποίο βασίζεται τόσο στα δεδομένα από τα οποίο έχει εξαχθεί ένα πρότυπο (extensional description) όσο και στην έννοια/ νόημα που αυτό αναπαριστά (intensional description).

Λέξεις κλειδιά πρότυπα, δέντρα απόφασης, συχνά στοιχειοσύνολα, κανόνες συσχέτισης, συστάδες, συσταδοποιήσεις.

2.1 Εισαγωγή

Η διαδικασία ΑΓΒΔ και το βήμα της Εξόρυξης Γνώσης προσφέρουν μία λύση στο πρόβλημα της υπερπληθώρας δεδομένων εξάγοντας έγκυρα, πρωτότυπα, πιθανόν χρήσιμα, και ευκόλως κατανοητά πρότυπα από τα δεδομένα [22]. Τα *πρότυπα* (patterns) αποτελούν συμπαγείς και περιεκτικές σε σημασιολογία αναπαραστάσεις των πρωτογενών δεδομένων [70] - συμπαγείς καθώς συνοψίζουν, σε κάποιο βαθμό, τις πληροφορίες που περιέχονται στα πρωτογενή δεδομένα, και πλούσια σε σημασιολογία καθώς αποκαλύπτουν νέα γνώση που βρίσκεται κρυμμένη στο μεγάλο όγκο των πρωτογενών δεδομένων.

Αρκετοί *τύποι προτύπων* έχουν προταθεί, εξαιτίας κυρίως της μεγάλης ανομοιογένειας των δεδομένων και των εφαρμογών Εξόρυξης Γνώσης, καθώς επίσης και εξαιτίας των διαφορετικών στόχων που κάθε τύπος προτύπου προσπαθεί να επιτύχει (δηλαδή, ποια χαρακτηριστικά των πρωτογενών δεδομένων προσπαθεί να τονίσει). Οι διαφορετικές εργασίες Εξόρυξης Γνώσης προσφέρουν διαφορετική κατανόηση των δεδομένων: τα συχνά στοιχειοσύνολα πιάνουν τις συσχετίσεις μεταξύ αντικειμένων, οι συστάδες αποκαλύπτουν φυσικές ομάδες στα δεδομένα, τα δέντρα απόφασης εντοπίζουν τα χαρακτηριστικά που προβλέπουν την τιμή ενός συγ-

κεκριμένου γνωρίσματος σε μελλοντικά, άγνωστα στιγμιότυπα του προβλήματος κ.ο.κ.

2.2 Αναπαράσταση προτύπων

Τα πρότυπα αποτελούν συμπαγείς και πλούσιες σε σημασιολογία αναπαραστάσεις των πρωτογενών δεδομένων. Συνεπώς, η αναπαράσταση ενός προτύπου μπορεί να είναι *αναλυτική* (extensional), με βάση τα δεδομένα από τα οποία εξάχθηκε ή *περιγραφική/διαισθητική* (intensional), με βάση το νόημα/έννοια που αναπαριστά. Ο *αναλυτικός τρόπος αναπαράστασης με βάση τα δεδομένα* στηρίζεται σε μία απαρίθμηση των δεδομένων από τα οποία εξήχθηκε το πρότυπο και συνεπώς είναι κοινός για τους διάφορους τύπους προτύπων. Από την άλλη, ο *περιγραφικός τρόπος αναπαράστασης με βάση το νόημα* που αναπαριστά το πρότυπο αποκαλύπτει πληροφορίες σχετικά με τη μορφή του προτύπου και τη σημασιολογία του, συνεπώς εξαρτάται από τον εκάστοτε τύπο προτύπων.

Όσον αφορά στην περιγραφική αναπαράσταση των προτύπων, ακολουθούμε τη λεγόμενη *ιδιότητα 2-συνιστωσών* των προτύπων (2-component property of patterns) που προτάθηκε στην [25]. Με βάση αυτή την ιδιότητα, πολλοί τύποι προτύπων μπορούν να περιγραφούν μέσω μιας δομικής συνιστώσας (structure component) και μιας ποσοτικής συνιστώσας (measure component). Η *δομική συνιστώσα* περιγράφει τη δομή των προτύπων που αποτελούν στιγμιότυπα του συγκεκριμένου τύπου προτύπων, π.χ., είναι το *head* και το *body* στην περίπτωση ενός κανόνα συσχέτισης. Η *ποσοτική συνιστώσα* συσχετίζει τα πρότυπα με τα πρωτογενή δεδομένα από τα οποία εξήχθησαν, π.χ., είναι η υποστήριξη (*support*) και η εμπιστοσύνη ($confidence$) στην περίπτωση ενός κανόνα συσχέτισης. Με άλλα λόγια, η δομική συνιστώσα περιγράφει το χώρο των προτύπων (pattern space), ενώ η ποσοτική συνιστώσα ποσοτικοποιεί το πόσο καλά ό χώρος των προτύπων περιγράφει το χώρο των πρωτογενών δεδομένων (data space).

Η ιδιότητα 2-συνιστωσών έχει επεκταθεί στα πλαίσια του έργου $PANDA$ [65], όπου παρουσιάζεται ένα *ενοποιημένο μοντέλο* για την αναπαράσταση των διαφόρων τύπων προτύπων. Το μοντέλο αυτό, εκτός από τη δομική και την ποσοτική συνιστώσα, περιλαμβάνει επίσης μία συνιστώσα πηγή (source component) που περιγράφει τα δεδομένα από τα οποία έγινε η εξαγωγή των προτύπων και μία συνιστώσα έκφρασης (expression component) που περιγράφει τη συσχέτιση μεταξύ του χώρου των δεδομένων και του χώρου των προτύπων (βλέπε Ενότητα 1.3.1).

Παρόμοιες ιδέες εμφανίζονται και στην [38], όπου οι συγγραφείς προτείνουν το μοντέλο των 3-Κόσμων. Στο μοντέλο αυτό, τα πρότυπα περιγράφονται ως περιορισμοί πάνω στο χώρο των γνωρισμάτων, ενώ διατηρούνται επίσης και οι συσχετίσεις με τα πραγματικά δεδομένα από τα οποία έγινε η εξαγωγή των προτύπων (βλέπε Ενότητα 1.3.1).

Ανακεφαλαιώνοντας, στα πλαίσια της συγκεκριμένης διατριβής υιοθετούμε την αναπαράσταση των προτύπων τόσο με βάση τα δεδομένα από τα οποία εξήχθησαν όσο και με βάση το νόημα/έννοια που αναπαριστούν. Όσον αφορά στην αναπαράσταση με βάση τα δεδομένα, τα πρότυπα περιγράφονται απαριθμώντας τα δεδομένα από τα οποία εξήχθησαν μέσω κάποιας τεχνικής Εξόρυξης Γνώσης. Όσον αφορά στην αναπαράσταση με βάση το νόημα/έννοια που αναπαριστούν, υιοθετούμε την λεγόμενη ιδιότητα 2-συνιστωσών, βάσει της οποίας ένα πρότυπο αποτελείται από μία δομική και από μία ποσοτική συνιστώσα.

2.3 Δέντρα απόφασης

Τα Δέντρα Απόφασης - ΔΑ (Decision Tree - DT) αποτελούν πολύ δημοφιλή *μοντέλα κατηγοριοποίησης/ταξινόμησης* εξαιτίας της διαισθητικής τους αναπαράστασης που τα καθιστά εύκολα κατανοητά από τους τελικούς χρήστες. Στην ενότητα αυτή παρουσιάζουμε μερικές βασικές έννοιες στα ΔΑ [50].

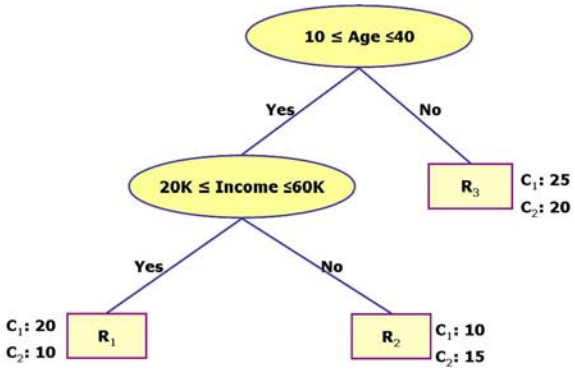
Έστω $A = \{A_1, A_2, ..., A_m\}$ είναι το σύνολο των γνωρισμάτων με βάση τα οποία θα γίνει η κατηγοριοποίηση (*γνωρίσματα πρόβλεψης*), όπου $dom(A_i)$ είναι το πεδίο ορισμού του γνωρίσματος A_i. Έστω C είναι το *γνώρισμα κλάση*, δηλαδή, το προς πρόβλεψη γνώρισμα, με πεδίο τιμών $dom(C) = \{C_1, C_2, ..., C_k\}$, όπου k είναι το πλήθος των κλάσεων. Η κατανομή πιθανότητας (probability distribution) των γνωρισμάτων πρόβλεψης, δηλαδή, $A = D(A_1) \times D(A_2) \times \cdots \times D(A_m)$, καλείται *κατανομή του χώρου των γνωρισμάτων* (attribute space distribution). Η από κοινού κατανομή πιθανότητας (joint probability distribution) των γνωρισμάτων πρόβλεψης και του γνωρίσματος κλάσης, $P = dom(A_1) \times dom(A_2) \times \cdots \times dom(A_m) \times dom(C)$, καλείται *κατανομή πιθανότητας του προβλήματος*.

Ο στόχος ενός δέντρου απόφασης είναι η εκμάθηση μίας συνάρτησης $f : dom(A_1) \times dom(A_2) \times \cdots \times dom(A_m) \rightarrow dom(C)$. Για το λόγο αυτό, χρησιμοποιείται ένα σύνολο από στιγμιότυπα του προβλήματος τα οποία προέρχονται από την κατανομή P και είναι γνωστά ως *στιγμιότυπα εκπαίδευσης* (training set) D. Ένα δέντρο απόφασης T που δημιουργείται μέσω του συνόλου στιγμιότυπων εκπαίδευσης D προσφέρει μία κατηγοριοποίηση των στιγμιότυπων του D στις κλάσεις C_j, $j = 1 \ldots k$ με βάση τις τιμές των γνωρισμάτων πρόβλεψης A_i, $i = 1 \ldots m$.

Τα γνωρίσματα πρόβλεψης μπορεί να είναι αριθμητικά (numerical), κατηγορηματικά (categorical) ή διατεταγμένα (ordinal). Το πεδίο ορισμού ενός *αριθμητικού γνωρίσματος* είναι ένα διατεταγμένο σύνολο (π.χ., ηλικία, εισόδημα), το πεδίο ορισμού ενός *κατηγορηματικού γνωρίσματος* είναι ένα πεπερασμένο σύνολο χωρίς κάποια φυσική κατηγοριοποίηση (π.χ., χρώμα, φύλλο, οικογενειακή κατάσταση), ενώ το πεδίο ορισμού ενός *διατεταγμένου γνωρίσματος* είναι ένα σύνολο από διατεταγμένες διακριτές τιμές χωρίς ωστόσο να υπάρχει γνώση σχετικά με τις απόλυτες διαφορές μεταξύ των τιμών αυτών (π.χ., κλίμακα προτιμήσεων, σοβαρότητα ενός τραυματισμού). Τα γνωρίσματα πρόβλεψης είναι κυρίως αριθμητικά.

Όσον αφορά στη δομή του, ένα ΔΑ αποτελείται από εσωτερικούς κόμβους και κόμβους - φύλλα. Ένας *εσωτερικός κόμβος* σχετίζεται με μία συνθήκη ελέγχου (splitting predicate) που καθορίζει έναν έλεγχο πάνω σε κάποιο γνώρισμα πρόβλεψης (π.χ., "$Age \geq 20$"). Κάθε διακλάδωση που ξεκινά από τον συγκεκριμένο κόμβο αντιστοιχεί σε μία από τις πιθανές τιμές αυτού του γνωρίσματος. Πιο συνηθισμένες είναι οι δυαδικές δηλώσεις της μορφής "Ναι" ή "Οχι". Ένας *κόμβος - φύλλο* περιγράφει την κλάση των στιγμιότυπων που ακολουθούν το μονοπάτι από τη ρίζα στο συγκεκριμένο κόμβο. Αν τα στιγμιότυπα ανήκουν σε παραπάνω από μία κλάσεις, η ετικέτα αυτή μπορεί να είναι η κλάση της πλειοψηφίας των στιγμιότυπων. Στη γενική περίπτωση, ένας κόμβος φύλλο μπορεί να συσχετίζεται σε κάποιο βαθμό με όλες τις κλάσεις του προβλήματος, ο βαθμός αυτός εξαρτάται από το ποσοστό των στιγμιότυπων που καταλήγουν στο φύλλο και ανήκουν στη συγκεκριμένη κλάση.

Στο Σχήμα 2.1 παρουσιάζουμε ένα παράδειγμα ενός δέντρου απόφασης, το οποίο αναφέρεται στο πρόβλημα της χορήγησης δανείων από μία τράπεζα. Υπάρχουν δύο γνωρίσματα πρόβλεψης: Age και $Income$, ενώ το γνώρισμα κλάσης C περιέχει δύο τιμές: $C = \{C_1, C_2\}$. Ένα μικρό τμήμα του συνόλου δεδομένων που χρησιμοποιήθηκε για τη δημιουργία του δέντρου απόφασης του Σχήματος 2.1



Σχήμα 2.1: Ένα παράδειγμα δέντρου απόφασης

παρουσιάζεται στον Πίνακα 2.1.

Instance	*Age*	*Salary*	*Class*
1	30	30K	C_1
2	35	10K	C_2
3	50	100K	C_1

Πίνακας 2.1: Ένα δείγμα του συνόλου δεδομένων εκπαίδευσης για το ΔΑ του Σχήματος 2.1

Αξιολόγηση ενός δέντρου απόφασης Όπως αναφέρθηκε ήδη, ένα ΔΑ δημιουργείται με βάση ένα σύνολο στιγμιότυπων εκπαίδευσης D που προέρχεται από την P, την από κοινού κατανομή πιθανότητας των γνωρισμάτων πρόβλεψης και του γνωρίσματος κλάσης. Ένα πλήρως ανεπτυγμένο δέντρο θα κατηγοριοποιούσε τέλεια το σύνολο στιγμιότυπων εκπαίδευσης. Ωστόσο, ένα ΔΑ δεν πρέπει μόνο να "ταιριάζει" στο σύνολο των στιγμιότυπων εκπαίδευσης, αλλά θα πρέπει επίσης να προβλέπει σωστά την κλάση άγνωστων στιγμιότυπων του προβλήματος - η ιδιότητα αυτή είναι γνωστή ως *ακρίβεια γενίκευσης* (generalization accuracy) του ΔΑ. Η *υπερπλήρωση* (overfitting) του συνόλου εκπαίδευσης είναι λάθος ιδιότητα για ένα δέντρο απόφασης, καθώς καλύπτει κάθε ιδιαιτερότητα των στιγμιότυπων αυτού του συνόλου και πολλές από τις ιδιαιτερότητες αυτές ενδέχεται να μην ξαναεμφανιστούν σε μελλοντικά στιγμιότυπα του προβλήματος. Η ακρίβεια γενίκευσης ενός ΔΑ αξιολογείται μέσω του *λάθους κατηγοριοποίησης* (miss-classification error - ME), το οποίο βασίζεται στο πλήθος των στιγμιότυπων για τα οποία το ΔΑ πρόβλεψε λάθος κλάση.

Ιδανικά, θα θέλαμε να γνωρίζουμε το ME του ταξινομητή f στην P, την κατανομή του προβλήματος. Ωστόσο, δεδομένου ότι δεν γνωρίζουμε την P (το μόνο που ξέρουμε είναι μερικά στιγμιότυπα από την κατανομή αυτή, τα οποία αποτελούν το σύνολο εκπαίδευσης), έχουν αναπτυχθεί διάφορες τεχνικές που υπολογίζουν το λάθος κατηγοριοποίηση του ταξινομητή f στην κατανομή P, δηλαδή υπολογίζουν το $ME(f, P)$. Η πιο κοινή προσέγγιση είναι αυτή του υπολογισμού μέσω ενός *holdout* συνόλου ελέγχου: το αρχικό σύνολο D των στιγμιότυπων του προβλήματος χωρίζεται σε δύο διακριτά σύνολα: το σύνολο εκπαίδευσης και το σύνολο ελέγχου. Το *σύνολο εκπαίδευσης* (training set) χρησιμοποιείται για

τη δημιουργία του ταξινομητή, ενώ το *σύνολο ελέγχου* (test set) χρησιμοποιείται για την αποτίμηση της απόδοσής του. Συνήθως, το 1/3 των στιγμιότυπων χρησιμοποιείται για τον έλεγχο και τα 2/3 για την εκπαίδευση. Άλλες δημοφιλείς τεχνικές σε αυτή την κατηγορία είναι οι re–substitution estimation και V–fold cross validation.

Μία λύση στο πρόβλημα της υπερπλήρωσης είναι το *κλάδεμα* (tree pruning). Ξεκινώντας από το τελευταίο επίπεδο του δέντρου, οι κόμβοι-παιδιά απομακρύνονται αν η απομάκρυνση αυτή επιφέρει αλλαγή στην ακρίβεια του δέντρου a φορές μικρότερη σε σχέση με την αλλαγή στην πολυπλοκότητα του δέντρου. Εξαιτίας του κλαδέματος, το δέντρο που προκύπτει μπορεί να μην προβλέπει τέλεια τα στιγμιότυπα του συνόλου εκπαίδευσης, το λάθος αυτό καλείται re-substitution error. Συνεπώς, ένα καλό ΔΑ θα πρέπει να ελαχιστοποιεί τόσο το re–substitution error (όσον αφορά στο σύνολο εκπαίδευσης) όσο και το λάθος κατηγοριοποίησης (όσον αφορά ένα ανεξάρτητο σύνολο εκπαίδευσης).

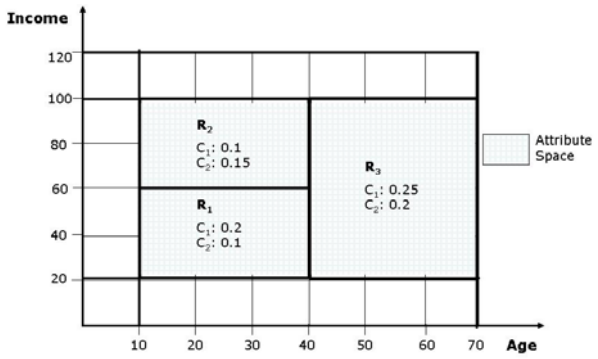
Τμηματοποίηση του χώρου γνωρισμάτων Ένα δέντρο απόφασης τμηματοποιεί το *χώρο γνωρισμάτων*, δηλαδή, το χώρο που ορίζεται από τα γνωρίσματα πρόβλεψης $D(A_1) \times D(A_2) \times \cdots \times D(A_m)$, σε ένα σύνολο διακριτών περιοχών έτσι ώστε κάθε περιοχή να είναι ομοιογενής, δηλαδή να περιέχει στιγμιότυπα που ανήκουν στην ίδια κλάση. Αυτό βέβαια αναφέρεται στην αναλυτική περίπτωση, καθώς μετά το κλάδεμα μία περιοχή μπορεί να περιέχει στιγμιότυπα από παραπάνω από μία κλάσεις του προβλήματος. Η γραμμή που χωρίζει δύο γειτονικές περιοχές που χαρακτηρίζονται από διαφορετικές κλάσεις καλείται *όριο απόφασης* (decision boundary). Τα όρια απόφασης είναι παράλληλα στους άξονες των γνωρισμάτων καθώς κάθε συνθήκη ελέγχου περιλαμβάνει ένα μόνο γνώρισμα. Συνεπώς, οι *περιοχές αποφάσεων* (decision regions) αποτελούν τετράγωνα παράλληλα στους άξονες των γνωρισμάτων [38].

Κάθε κόμβος φύλλο του ΔΑ αντιστοιχεί σε μία περιοχή του R. Μία *περιοχή* μπορεί να περιγραφεί αναλυτικά (extensionally) με βάση τα στιγμιότυπα που καταλήγουν στον αντίστοιχο κόμβο-φύλλο. Μία περιοχή μπορεί επίσης να περιγραφεί διαισθητικά/ περιγραφικά (intensionally) με βάση το μονοπάτι του δέντρου που ξεκινάει από τη ρίζα του δέντρου και καταλήγει στον αντίστοιχο κόμβο-φύλλο. Πιο συγκεκριμένα, η διαισθητική περιγραφή μιας περιοχής αποτελείται από μία δομική συνιστώσα και από μία ποσοτική συνιστώσα. Η *δομική συνιστώσα μιας περιοχής* (structure component) αποτελείται από την συνένωση των συνθηκών ελέγχου κατά μήκος του αντίστοιχου μονοπατιού, ενώ η *ποσοτική συνιστώσα μια περιοχής* (measure component) αποτελείται από την κατανομή των στιγμιότυπων της περιοχής στις διάφορες κλάσεις του προβλήματος.

Η τμηματοποίηση του δέντρου του Σχήματος 2.1 απεικονίζεται στο Σχήμα 2.2.

Ας δούμε τώρα πως η περιοχή R_1 του ΔΑ (βλέπε Σχήμα 2.2) μπορεί να περιγραφεί μέσω του αναλυτικού-περιγραφικού τρόπου αναπαράστασης (extensional-intensional representation schema). Η δομική συνιστώσα είναι: $R_1.struct = (10 \leq Age \leq 40) \cap (20K \leq Income \leq 60K)$, ενώ η ποσοτική συνιστώσα είναι: $R_1.meas = \{(C_1 : 20\%), (C_2 : 10\%)\}$· οι δύο αυτές συνιστώσες αποτελούν την περιγραφική αναπαράσταση της περιοχής R_1. Η αναλυτική περιγραφή της R_1 αποτελείται από το υποσύνολο των στιγμιότυπων του D που καταλήγουν στην R_1.

Σημειώστε ότι ο χώρος γνωρισμάτων καθορίζεται με βάση τα γνωρίσματα πρόβλεψης του προβλήματος, συνεπώς είναι κοινός για όλα τα ΔΑ που αναφέρονται στο



Σχήμα 2.2: Η τμηματοποίηση του χώρου γνωρισμάτων για το ΔΑ της Εικόνας 2.1

ίδιο πρόβλημα κατηγοριοποίησης. Αυτό που διαφοροποιεί τα διάφορα ΔΑ είναι η τμηματοποίηση που επιφέρουν πάνω στο χώρο των γνωρισμάτων δηλαδή, ποιες είναι οι περιοχές που δημιουργούνται.

Μέχρι στιγμής, εστιάσαμε στον τρόπο αναπαράστασης μίας συγκεκριμένης περιοχής ενός ΔΑ με βάση τον αναλυτικό-περιγραφικό τρόπο αναπαράστασης. Ένα ΔΑ αποτελείται από ένα *σύνολο από περιοχές*, όπου κάθε περιοχή μπορεί να περιγραφεί με βάση τον αναλυτικό-περιγραφικό τρόπο αναπαράστασης που αναφέραμε πριν.

Περισσότερες λεπτομέρειες σχετικά με την τμηματοποίηση που επιτυγχάνει ένα ΔΑ στο χώρο των γνωρισμάτων υπάρχουν στο Κεφάλαιο 5.

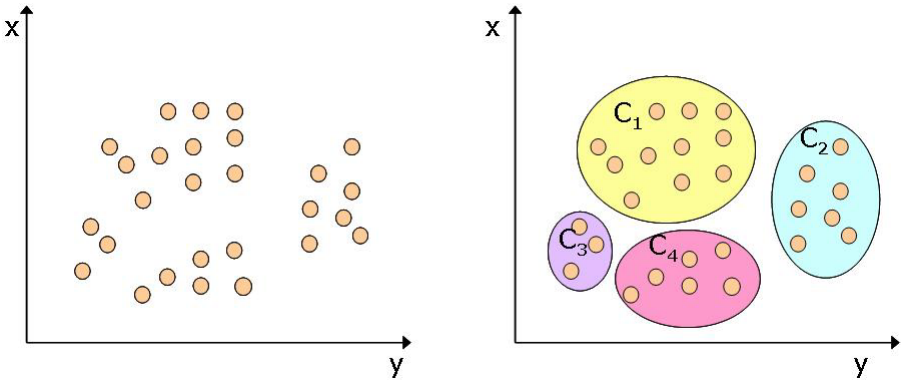
2.4 Συστάδες και συσταδοποιήσεις

Η συσταδοποίηση (clustering) είναι η μη-επιβλεπόμενη κατηγοριοποίηση των δεδομένων σε φυσικές ομάδες, καλούμενες *συστάδες* (clusters), προκειμένου αντικείμενα που ανήκουν στην ίδια ομάδα να είναι πιο όμοια μεταξύ τους σε σχέση με αντικείμενα που ανήκουν σε διαφορετικές ομάδες [36]. Ο όρος *μη-επιβλεπόμενη* αναφέρεται στο γεγονός ότι δεν υπάρχει εκ των προτέρων γνώση σχετικά με την τμηματοποίηση των δεδομένων. Πιο τυπικά, μπορούμε να πούμε ότι μία συσταδοποίηση ζ είναι η τμηματοποίηση ενός συνόλου δεδομένων D στις συστάδες $C_1, C_2, \ldots, C_k$ έτσι ώστε $C_i \cap C_j = \emptyset$ και $\cup_{j=1}^{k} C_j = D$. Αυτός ο ορισμός ισχύει για την *αυστηρή συσταδοποίηση* (hard clustering), όπου ένα στιγμιότυπο ανατίθεται σε ακριβώς μία συστάδα δημιουργώντας έτσι μία ξεκάθαρη τμηματοποίηση του συνόλου δεδομένων. Ένας πιο χαλαρός ορισμός για την τμηματοποίηση είναι αυτός της *χαλαρής συσταδοποίησης* (soft clustering) όπου ένα στιγμιότυπο επιτρέπεται να ανήκει σε περισσότερες από μία συστάδες με βάση κάποιο βαθμό συμμετοχής.

Οι αλγόριθμοι συσταδοποίησης βασίζονται σε κάποια *συνάρτηση απόστασης* (distance function) που καθορίζει σε ποια συστάδα πρέπει να ανατεθεί ένα στιγμιότυπο. Μία ευρέως χρησιμοποιούμενη συνάρτηση απόστασης είναι η Ευκλείδεια απόσταση. Υπάρχει επίσης και μία συνάρτηση αξιολόγησης (evaluation function) που αξιολογεί την ποιότητα της συσταδοποίησης. Συνήθως μία τέτοια συνάρτηση αποσκοπεί στην ελαχιστοποίηση της απόστασης κάθε στιγμιότυπου από το κέντρο της συστάδας στην οποία ανατέθηκε μέσω της διαδικασίας της συσταδοποίησης.

Αλγόριθμοι συσταδοποίησης Το πρόβλημα της συσταδοποίησης έχει μελετηθεί εκτεταμένα σε πολλούς τομείς συμπεριλαμβανόμενης και της Εξόρυξης Γνώσης. Για το λόγο αυτό ένας μεγάλος αριθμών αλγορίθμων συσταδοποίησης έχει προταθεί (βλέπε [36] για μία εκτενή ανασκόπηση). Παρόλο που οι διάφοροι αλγόριθμοι χρησιμοποιούν διάφορους ορισμούς για την έννοια της συστάδας, μπορούν να ταξινομηθούν ως ακολούθως [30]:

1. Οι *διαμεριστικές μέθοδοι* (partitioning methods) που τμηματοποιούν το σύνολο δεδομένων σε k ομάδες/ συστάδες όπου κάθε μία αναπαρίσταται μέσω ενός κεντροειδούς (centroid) όπως στον $k-means$ ή μέσω ενός *medoid* όπως στον $k-medoids$. Ο αριθμός των συστάδων, k, ορίζεται από τον χρήστη. Ένα παράδειγμα του αλγορίθμου $k-means$ παρουσιάζεται στο Σχήμα 2.3.

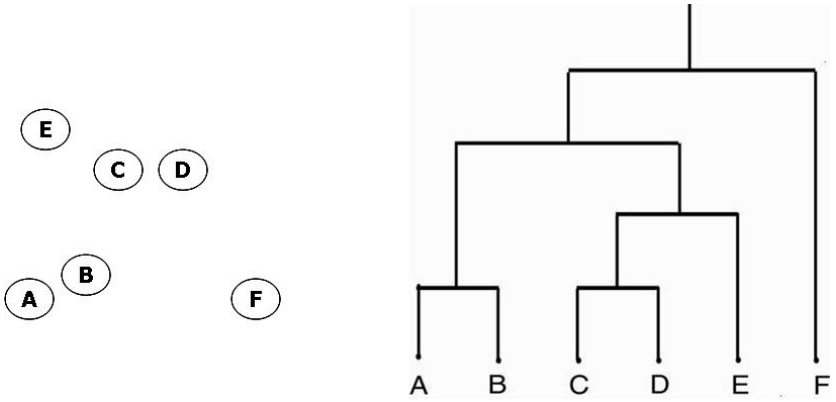


Σχήμα 2.3: Παράδειγμα ενός μικρού συνόλου δεδομένων (αριστερά) και η αντίστοιχη $k-means$ συσταδοποίηση για $k=4$ (δεξιά)

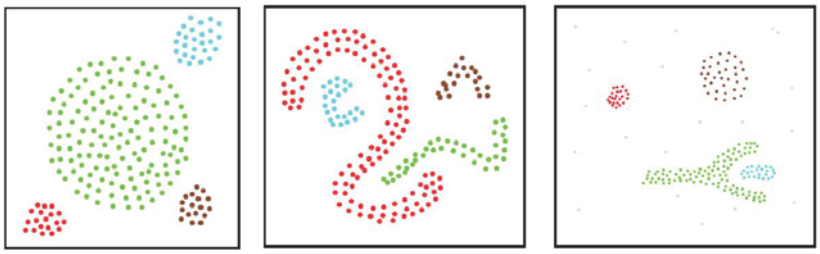
2. Οι *ιεραρχικές μέθοδοι* (hierarchical methods) που δημιουργούν μία ιεραρχική διάσπαση του συνόλου δεδομένων που καλείται δενδρόγραμμα. Μία τέτοια διάσπαση μπορεί να δημιουργηθεί είτε με $bottom-up$ τρόπο είτε με $top-down$ τρόπο, δημιουργώντας *agglomerative hierarchical methods* ή *divisive hierarchical methods*, αντίστοιχα. Και στις δύο περιπτώσεις, χρειάζεται να έχει οριστεί μία συνάρτηση απόστασης μεταξύ των συστάδων. Διαφορετικές συναρτήσεις απόστασης μπορούν να χρησιμοποιηθούν όπως οι single linkage, complete linkage, average linkage ή η απόσταση των κεντροειδών (centroids distance) [30]. Ένα παράδειγμα δενδρογράμματος παρουσιάζεται στο Σχήμα 2.4.

3. Οι *μέθοδοι με βάση την πυκνότητα* (density based methods) που συνεχίζουν να αναπτύσσουν μία συστάδα όσο η πυκνότητα (δηλαδή, το πλήθος των αντικειμένων στην γειτονιά του) υπερβαίνει ένα κατώφλι. Στην κατηγορία αυτή ανήκει ο αλγόριθμος $DBSCAN$, μερικά παραδείγματα του οποίου παρουσιάζονται στο Σχήμα 2.5.

4. Οι *βασιζόμενες στο πλέγμα μέθοδοι συσταδοποίησης* (grid based methods) που χωρίζουν το χώρο σε ένα πεπερασμένο αριθμό κελιών που ορίζουν μία δομή πλέγματος. Στην κατηγορία αυτή ανήκουν οι αλγόριθμοι $STING$ και $CLIQUE$.



Σχήμα 2.4: Παράδειγμα ενός μικρού συνόλου δεδομένων (αριστερά) και το αντίστοιχο δενδρόγραμμα (δεξιά)



Σχήμα 2.5: Παραδείγματα του αλγορίθμου $DBScan$

5. Οι *μέθοδοι συσταδοποίησης με βάση κάποιο μοντέλο* (model based methods) που υποθέτουν ένα μοντέλο για κάθε συστάδα και βρίσκουν το καλύτερο ταίριασμα των δεδομένων στο δοθέν μοντέλο. Οι στατιστικές προσεγγίσεις, όπως ο αλγόριθμος $COBWEB$, EM, και οι προσεγγίσεις με βάση τα νευρωνικά δίκτυα ανήκουν στην κατηγορία αυτή.

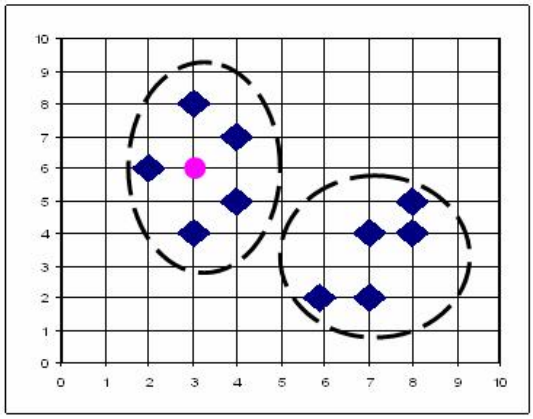
Αναπαράσταση συστάδων Η *αναλυτική περιγραφή μιας συστάδας με βάση τα δεδομένα* είναι απλή· αποτελείται από μία αναλυτική απαρίθμηση των δεδομένων που πέφτουν στα όρια της συστάδας. Συνεπώς, η αναλυτική περιγραφή είναι κοινή για τους διάφορους τύπους συστάδων.

Ωστόσο, η *διαισθητική περιγραφή μίας συστάδας με βάση το νόημα που αναπαριστά* εξαρτάται από τον συγκεκριμένο τύπο συστάδας. Για παράδειγμα, στην περίπτωση μιας διαμεριστικής συσταδοποίησης μία συστάδα μπορεί να περιγραφεί μέσω του κεντροειδούς της όπως στον $k-means$ ή μέσω του $centroid$ όπως στον $k-medoids$. Στην περίπτωση των αλγορίθμων συσταδοποίησης με βάση την πυκνότητα, η συστάδα μπορεί να περιγραφεί με βάση κάποια συνάρτηση κατανομής πυκνότητας πιθανότητας. Να σημειώσουμε ωστόσο πως υπάρχουν αλγόριθμοι όπως οι ιεραρχικοί για τους οποίους δεν μπορεί να οριστεί κάποια διαισθητική περιγραφή για τις συστάδες, Στην περίπτωση αυτή, μία συστάδα μπορεί να αναπαρασταθεί αναλυτικά με βάση τα αντικείμενα που την απαρτίζουν.

Όσον αφορά στην *ποσοτική συνιστώσα* (measure component) μιας συστάδας,

υπάρχουν διάφορα εναλλακτικά μέτρα που μπορούν να χρησιμοποιηθούν, όπως για παράδειγμα η υποστήριξη της συστάδας (δηλαδή, το ποσοστό των στιγμιότυπων που ανατίθεται στη συστάδα) ή η $intra - cluster$ απόσταση (δηλαδή, η μέση απόσταση μεταξύ των αντικειμένων της συστάδας) ή η μέση απόσταση των μελών της συστάδας από το κέντρο της συστάδας.

Προκειμένου να κάνουμε πιο κατανοητό τον τρόπο αναπαράστασης μίας συστάδας, παρουσιάζουμε στο Σχήμα 2.6 ένα παράδειγμα συσταδοποίησης μέσω του $k - means$. Ας πάρουμε για παράδειγμα την αριστερή συστάδα του παραπάνω σχήματος: Μπορούμε να περιγράψουμε τη συστάδα αυτή αναλυτικά με βάση τα δεδομένα που περιέχει, δηλαδή, {(3,4), (3,8), (4,5), (4,7), (2,6)}. Επίσης, μπορούμε να την αναπαραστήσουμε και περιγραφικά με βάση το κέντρο της $(3.2, 6)$ και το πλήθος των μελών της συστάδας, δηλαδή, 5.



Σχήμα 2.6: Παράδειγμα αναλυτικής (με βάση τα δεδομένα) και περιγραφικής (με βάση το νόημα) αναπαράστασης συστάδας

2.5 Συχνά στοιχειοσύνολα και κανόνες συσχέτισης

Το *πρόβλημα της εξόρυξης συχνών στοιχειοσυνόλων* (Frequent Itemset Mining -FIM) αποτελεί ένα από τα βασικά προβλήματα στον τομέα της Εξόρυξης Γνώσης με πολλές εφαρμογές όπως οι κανόνες συσχέτισης και τα ακολουθιακά πρότυπα.

Για τον ορισμό του προβλήματος, ακολουθούμε την εργασία [4]: Έστω I είναι ένα πεπερασμένο σύνολο από διακριτά στοιχεία και D μία βάση δεδομένων συναλλαγών όπου κάθε συναλλαγή T περιέχει ένα σύνολο από στοιχεία, $T \subseteq I$. Ένα παράδειγμα μιας βάσης δεδομένων συναλλαγών παρουσιάζεται στον Πίνακα 2.2.

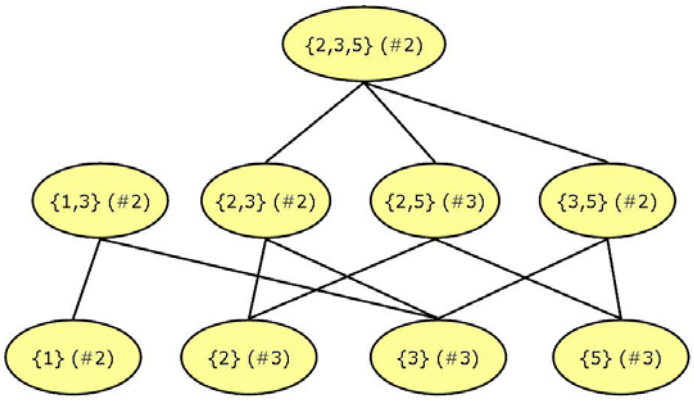
Ένα *στοιχειοσύνολο* (itemset) X είναι ένα μη κενό, λεξικογραφικά διατεταγμένο σύνολο στοιχείων, $X \subseteq I$. Αν το X αποτελείται από k στοιχεία, καλείται $k - itemset$. Η συχνότητα του X στο D ισούται με το πλήθος των συναλλαγών του D που περιέχουν το X, δηλαδή, $fr_D(X) = |\{T \in D : X \subseteq T\}|$. Το ποσοστό των συναλλαγών του D που περιέχει το X, καλείται *υποστήριξη* (support) του X στην D, $supp_D(X) = \frac{fr_D(X)}{|D|}$. Ένα *συχνό στοιχειοσύνολο* (frequent itemset) είναι ένα στοιχειοσύνολο με υποστήριξη μεγαλύτερη ή ίση ενός ελάχιστου κατωφλίου σ,

Transaction ID	Transaction Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Πίνακας 2.2: Ένα παράδειγμα μίας βάσης δεδομένων συναλλαγών

που καλείται $minSupport$, δηλαδή, $supp_D(X) \geq \sigma$. Δύο συχνά στοιχειοσύνολα ανήκουν στην ίδια κλάση ισοδυναμίας F_k αν μοιράζονται ένα κοινό πρόθεμα μήκους $k-1$. Ο στόχος του προβλήματος της εξόρυξης συχνών στοιχειοσυνόλων είναι η εξαγωγή όλων των στοιχειοσυνόλων X από τη βάση δεδομένων D που είναι συχνά με βάση ένα κατώφλι ελάχιστης υποστήριξης σ. Έστω $F_\sigma(D)$ είναι το σύνολο των συχνών στοιχειοσυνόλων που έχει εξαχθεί από τη βάση δεδομένων D κάτω από ένα κατώφλι ελάχιστης υποστήριξης σ.

Το σύνολο των συχνών στοιχειοσυνόλων (FI) σχηματίζει το *πλέγμα των συχνών στοιχειοσυνόλων* L στο οποίο ισχύει η λεγόμενη *apriori* ιδιότητα: ένα στοιχειοσύνολο είναι συχνό αν όλα τα υποσύνολά του είναι επίσης συχνά. Για τη βάση δεδομένων του Πίνακα 2.2 και για $minSupport = 2$, το πλέγμα που προκύπτει απεικονίζεται στο Σχήμα 2.7.



Σχήμα 2.7: Ένα παράδειγμα πλέγματος συχνών στοιχειοσυνόλων

Αναπαράσταση συχνών στοιχειοσυνόλων Η *αναλυτική περιγραφή ενός συχνού στοιχειοσυνόλου* περιλαμβάνει μια απαρίθμηση των συναλλαγών που το υποστηρίζουν. Η *διαισθητική περιγραφή ενός στοιχειοσυνόλου* αποτελείται από το σύνολο των στοιχείων που το απαρτίζουν (αυτό αποτελεί τη *δομική συνιστώσα*) και από την υποστήριξη (αυτό αποτελεί την *ποσοτική συνιστώσα*. Ας θεωρήσουμε για παράδειγμα το στοιχειοσύνολο $(\{1,3\}, \#2)$ του Σχήματος 2.7. Η αναλυτική του περιγραφή αποτελείται από τις συναλλαγές που το υποστηρίζουν $\{100, 300\}$ (βλέπε Πίνακας 2.2). Όσον αφορά στη διαισθητική του περιγραφή, η δομική συνιστώσα αποτελείται από τα στοιχεία $\{1,3\}$, ενώ η ποσοτική συνιστώσα αποτελείται από την υποστήριξή του που ισούται με 2.

Εξαγωγή κανόνων συσχέτισης (Association Rules Mining) Το *πρόβλημα της εξαγωγής κανόνων συσχέτισης* [4] ορίζεται ως εξής: Έστω D είναι μία βάση δεδομένων συναλλαγών, όπου κάθε συναλλαγή αποτελείται από ένα σύνολο διακριτών στοιχείων I, τα λεγόμενα στοιχειοσύνολα. Ένας *κανόνας συσχέτισης* είναί μία έκφραση της μορφής $X \rightarrow Y$, όπου $X \subseteq I$, $Y \subseteq I$ και $X \cap Y = \emptyset$ (X και Y είναι στοιχειοσύνολα). Ο κανόνας σχετίζεται με μία *υποστήριξη* s και μία *εμπιστοσύνη* c. Ένας κανόνας $X \rightarrow Y$ έχει υποστήριξη s, αν $s\%$ των συναλλαγών του D περιέχουν το $X \cup Y$, ενώ έχει εμπιστοσύνη c, αν $c\%$ των συναλλαγών του D που περιέχουν το X περιέχουν και το Y. Ένας κανόνας καλείται *ενδιαφέρον* ή *ισχυρός* αν η υποστήριξη και η εμπιστοσύνη του υπερβαίνουν κάποια κατώφλια που ορίζει ο χρήστης.

Το πρόβλημα της εξαγωγής κανόνων συσχέτισης απαρτίζεται από δύο βήματα: Αρχικά εξάγεται το σύνολο των συχνών στοιχειοσυνόλων, και χρησιμοποιείται μετά ως είσοδος στο δεύτερο βήμα για την εξαγωγή των κανόνων συσχέτισης. Οι κανόνες συσχέτισης παρέχουν παραπάνω πληροφορία σε σχέση με τα στοιχειοσύνολα, καθώς περιγράφουν κατά πόσο η εμφάνιση ενός συνόλου από στοιχειοσύνολα συνεπάγεται κάποιο άλλο σύνολο.

Ας επιστρέψουμε τώρα στο θέμα της αναλυτικής - διαισθητικής περιγραφής των κανόνων συσχέτισης: η *αναλυτική περιγραφή ενός κανόνα συσχέτισης* αποτελείται από μία απαρίθμηση των συναλλαγών που συνεισφέρουν στη δημιουργία του. Για παράδειγμα, η αναλυτική περιγραφή του κανόνα $2 \rightarrow 5$ είναι το σύνολο των στιγμιότυπων $\{200, 400\}$ του Πίνακα 2.2. Η *διαισθητική περιγραφή ενός κανόνα συσχέτισης* αποτελείται από το *head* και το *body* που συνθέτουν τη δομική συνιστώσα και από την υποστήριξη και την εμπιστοσύνη που συνθέτουν την ποσοτική συνιστώσα. Στο παράδειγμά μας, $head$=$\{2\}$, $body$=$\{5\}$, $support = 50\%$ και $confidence = 100\%$.

2.6 Σύνοψη

Στο κεφάλαιο αυτό, παρουσιάσαμε μερικούς από τους πιο δημοφιλείς τύπους προτύπων, συγκεκριμένα τα δέντρα απόφασης, τις συστάδες/συσταδοποιήσεις και τα συχνά στοιχειοσύνολα/κανόνες συσχέτισης. Για την περιγραφή των προτύπων, χρησιμοποιήσαμε τόσο την αναλυτική τους περιγραφή (extensional description) με βάση τα δεδομένα από τα οποία εξήχθησαν όσο και τη διαισθητική τους περιγραφή (intensional description) με βάση το νόημα/ έννοια που περιγράφουν. Να τονίσουμε πως η αναλυτική περιγραφή είναι κοινή για όλους τους τύπους προτύπων, καθώς αποτελεί στην ουσία μία απαρίθμηση των στιγμιότυπων του συνόλου δεδομένων που στηρίζουν το εκάστοτε πρότυπο. Από την άλλη, η διαισθητική περιγραφή (ή περιγραφική αναπαράσταση) εξαρτάται από τον τύπο προτύπων, π.χ., ένα δέντρο απόφασης περιγράφεται διαφορετικά από ένα συχνό στοιχειοσύνολο.

Ο ορισμός ενός προτύπου είναι πλήρης αν είναι διαθέσιμες τόσο η αναλυτική όσο και η διαισθητική του περιγραφή. Ωστόσο, δεν μπορούμε πάντα να έχουμε και τις δύο περιγραφές, για λόγους π.χ., ιδιωτικότητας ή αποδοτικότητας (βλέπε Κεφάλαιο 1). Στη διατριβή αυτή, μελετάμε το θέμα της σύγκρισης προτύπων εκμεταλλευόμενοι κυρίως τη διαισθητική περιγραφή των προτύπων. Υπάρχουν, ωστόσο, περιπτώσεις που χρησιμοποιούμε επίσης και την αναλυτική τους περιγραφή, π.χ., στην παρακολούθηση συστάδων γενικού τύπου (ιεραρχικές, διαμεριστικές, με βάση την πυκνότητα). Παρόλο που, η πλήρης περιγραφή ενός προτύπου γίνεται μέσου του αναλυτικού-περιγραφικού σχήματος, υπάρχουν ωστόσο περιπτώσεις που

και η διαισθητική περιγραφή από μόνη της αποτελεί μία πολύ καλή περιγραφή του προτύπου.

Κεφάλαιο 3

Σύγκριση Προτύπων Αυθαίρετης Πολυπλοκότητας - Το Πλαίσιο *PANDA*

Σε αυτό το κεφάλαιο, παρουσιάζουμε το *PANDA* ένα γενικό και ευέλικτο πλαίσιο για αποτίμηση της ανομοιότητας μεταξύ προτύπων αυθαίρετης πολυπλοκότητας. Το *PANDA* χειρίζεται τόσο απλά όσο και σύνθετα πρότυπα, που ορίζονται πάνω σε πρωτογενή δεδομένα και σε άλλα πρότυπα, αντίστοιχα. Το όνομα *PANDA* αντιστοιχεί στο Patterns for Next Generation Database Systems, ένα ακρωνύμιο που χρησιμοποιήθηκε στο IST-2001-33058 ερευνητικό έργο της Ευρωπαϊκής Ένωσης, το οποίο εισάγει και μελετά την έννοια των Συστημάτων Διαχείρισης Βάσεων Προτύπων (Pattern Base Management Systems -PBMS) [64].

Το κεφάλαιο είναι οργανωμένο ως εξής: Στην Ενότητα 3.1, παρουσιάζονται τα κίνητρα της μελέτης και οι απαιτήσεις που θα πρέπει να ικανοποιεί ένα γενικό πλαίσιο αποτίμησης ανομοιότητας μεταξύ προτύπων. Το μοντέλο αναπαράστασης προτύπων περιγράφεται στην Ενότητα 3.2, ενώ η διαδικασία αποτίμησης της ανομοιότητας περιγράφεται στην Ενότητα 3.3. Στην Ενότητα 3.4, περιγράφουμε κάποιες λεπτομέρειες υλοποίησης, ενώ στην Ενότητα 3.5, εφαρμόζουμε το *PANDA* σε συγκεκριμένα προβλήματα σύγκρισης. Η σχετική βιβλιογραφία παρουσιάζεται στην Ενότητα 3.6. Στην Ενότητα 3.7 συνοψίζουμε τις βασικές ιδέες του *PANDA* ενώ στην Ενότητα 3.8, περιγράφουμε πιθανές βελτιώσεις και ανοιχτά θέματα.

Λέξεις κλειδιά σύγκριση προτύπων, πλαίσιο σύγκρισης προτύπων, απλά πρότυπα, σύνθετα πρότυπα.

3.1 Κίνητρα μελέτης και απαιτήσεις

Σε αυτή την ενότητα, θα παρουσιάσουμε μερικά επεξηγηματικά παραδείγματα τα οποία, σε συνδυασμό με τα παραδείγματα/εφαρμογές που έχουμε ήδη παρουσιάσει στην Ενότητα 1.4, μας ωθούν στη δημιουργία ενός γενικού και ευέλικτου πλαισίου σύγκρισης προτύπων. Τα παραδείγματα αυτά δεν υποστηρίζονται πλήρως από τις

υπάρχουσες τεχνικές, γεγονός που θα γίνει περισσότερο εμφανές στην Ενότητα 3.6 όπου παρουσιάζεται η σχετική βιβλιογραφία.

Παράδειγμα 1 *Φανταστείτε μία τηλεπικοινωνιακή εταιρία που παρέχει ένα πακέτο νέας γενιάς υπηρεσιών με βάση τα διαφορετικά προφίλ των πελατών της. Φανταστείτε επίσης ότι το στέλεχος που λαμβάνει αποφάσεις στην εταιρία ζητά μία μηνιαία αναφορά που θα παρουσιάζει την χρησιμοποίηση του πακέτου αυτού από τους πελάτες - χρήστες.*

□

Μία τέτοια αναφορά θα ήταν πολύ πιο ευανάγνωστη και χρήσιμη για το συγκεκριμένο στέλεχος, αν συνοδευόταν και από μία μηνιαία σύγκριση της κατηγοριοποίησης των προφίλ των πελατών - χρηστών της υπηρεσίας (Μια τέτοια κατηγοριοποίηση θα μπορούσε π.χ., να απεικονιστεί μέσω μοντέλων δέντρων απόφασης).

Παράδειγμα 2 *Ας σκεφτούμε τον μάνατζερ μίας αλυσίδας σουπερμάρκετ ο οποίος θέλει να αναλύσει τις τάσεις στις πωλήσεις κάθε καταστήματος της αλυσίδας. Συγκεκριμένα, ο μάνατζερ ενδιαφέρεται να μάθει αν υπάρχει κάποιο κατάστημα στο οποίο οι πωλήσεις διαφέρουν σημαντικά από τις πωλήσεις των άλλων καταστημάτων.*

□

Αυτή η ανάλυση θα μπορούσε να πραγματοποιηθεί με πολλούς διαφορετικούς τρόπους: ι) μελετώντας τις πωλήσεις μεμονωμένων προϊόντων, ιι) μελετώντας τα καλάθια αγορών των πελατών, ιιι) μελετώντας τα προϊόντα που χαρακτηρίζουν κάθε ομάδα πελατών.

Παράδειγμα 3 *Μία εφαρμογή Εξόρυξης Γνώσης από χωρικά δεδομένα αναλύει τις συσχετίσεις μεταξύ της πυκνότητας του πληθυσμού σε μία πόλη και του αριθμού των αυτοκινητιστικών ατυχημάτων που συμβαίνουν στην πόλη αυτή. Για λόγους ιδιωτικότητας, τα πρωτογενή δεδομένα δεν είναι διαθέσιμα. Αντίθετα, έχουμε στη διάθεσή μας μόνο τις κατανομές του πληθυσμού και των αυτοκινητιστικών ατυχημάτων στις περιοχές της πόλης.*

□

Η συσχέτιση που προκύπτει είναι υψηλή αν ληφθούν υπόψη οι (χωρικές) συσχετίσεις μεταξύ των γειτονικών περιοχών, ενώ αντίθετα, η συσχέτιση είναι χαμηλή όταν οι κατανομές συγκρίνονται ανά περιοχή και αγνοούνται τυχόν συσχετίσεις μεταξύ των περιοχών.

Παράδειγμα 4 *Ο κατασκευαστής ενός συστήματος εντοπισμού αντιγραφών πρέπει να πειραματιστεί με διαφορετικές τεχνικές προκειμένου να επιλέξει την πιο αποτελεσματική τεχνική για τη σύγκριση πολυμεσικών εγγράφων, τα οποία αναπαρίστανται μέσω ενός συνόλου γνωρισμάτων (π.χ., μία λίστα από λέξεις κλειδιά αν πρόκειται για κείμενα, την κατανομή των χρωμάτων αν πρόκειται για εικόνες, κτλπ.).*

□

Για το σκοπό αυτό, ο προγραμματιστής θα πρέπει να πειραματιστεί με διάφορες μεθόδους που λαμβάνουν υπόψη αυτά τα γνωρίσματα.

Παράδειγμα 5 *Ένας δικτυακός τόπος χαρακτηρίζεται από μεγάλη ποικιλία περιεχομένου καθώς καλύπτει μία μεγάλη γκάμα από θεματικές ενότητες. Αυτές οι θεματικές ενότητες είναι συνήθως προκαθορισμένες και οι χρήστες απλά προσθέτουν άρθρα σε κάθε ενότητα. Στην πράξη, οι θεματικές ενότητες μπορούν να είναι επικαλυπτόμενες ή κάποια άρθρα μπορούν από τη φύση τους να ενταχθούν σε παραπάνω από μία ενότητες. Ας θεωρήσουμε για παράδειγμα τις ενότητες "Βάσεις Δεδομένων" και "Εξόρυξη Γνώσης " και ένα άρθρο σχετικά με "Εξόρυξη γνώσης από μεγάλες Βάσεις Δεδομένων". Σε αυτή την περίπτωση, δεν είναι σαφές για τον χρήστη σε ποια ενότητα θα πρέπει να προσαρτήσει το άρθρο του και έτσι, μπορεί να το αναθέσει τυχαία σε μία από τις δύο ενότητες ή και στις δύο μαζί. Ως αποτέλεσμα, ο δικτυακός τόπος μπορεί να περιέχει ενότητες που παρουσιάζουν επικαλύψεις ως προς το περιεχόμενό τους.*

□

Για καλύτερη εξυπηρέτηση των αναγκών των χρηστών (π.χ., αναζήτηση, πλοήγηση), ο ιδιοκτήτης του δικτυακού τόπο θα μπορούσε να οργανώσει το περιεχόμενο του τόπου έτσι ώστε οι ετικέτες/τίτλοι των ενοτήτων να είναι ενδεικτικές των άρθρων που περιέχουν. Ως πρώτο βήμα προς αυτή την κατεύθυνση, οι παρόμοιες ενότητες θα πρέπει να αναγνωριστούν. Στη συνέχεια, ο ιδιοκτήτης θα πρεπε να σκεφτεί κατά πόσο, για παράδειγμα, δύο ενότητες μπορούν να συγχωνευτούν σε μία. Μία άλλη λύση θα μπορούσε να είναι η επανα-κατανομή των άρθρων στις διάφορες ενότητες, έτσι ώστε οι ενότητες να είναι πιο ομοιογενείς όσον αφορά στο περιεχόμενό τους και πιο ακριβείς όσον αφορά στις ετικέτες/ τίτλους τους.

Τα παραπάνω παραδείγματα υποδεικνύουν ότι ένα πλαίσιο για σύγκριση προτύπων θα πρέπει να ικανοποιεί τις παρακάτω βασικές απαιτήσεις:

Γενική εφαρμοσιμότητα: Το πλαίσιο θα πρέπει να μπορεί να εφαρμοστεί σε αυθαίρετους τύπους προτύπων, όπως κανόνες συσχέτισης, ιστογράμματα και γράφους. Την ίδια στιγμή, δεν θα πρέπει να περιορίζει την πολυπλοκότητα των υπό θεώρηση τύπων προτύπων. Το Παράδειγμα 5 που αναφέρεται στην σύγκριση δικτυακών τόπων αίρει μία τέτοια απαίτηση καθώς ένας δικτυακός τόπος αποτελεί μία σύνθετη δομή που αποτελείται από θεματικές ενότητες και άρθρα.

Ευελιξία: Το πλαίσιο θα πρέπει να επιτρέπει τον ορισμό εναλλακτικών συναρτήσεων ανομοιότητας, ακόμα και για τον ίδιο τύπο προτύπων. Πράγματι, προσωπικές προτιμήσεις των χρηστών και συγκεκριμένοι περιορισμοί όσον αφορά την εκτίμηση της ανομοιότητας θα πρέπει να μπορούν να ενσωματωθούν εύκολα στο πλαίσιο. Αυτή η απαίτηση περιγράφεται στο Παράδειγμα 4, όπου ο χρήστης πρέπει να πειραματιστεί με διάφορες παραμέτρους έτσι ώστε να αποφασίσει τις ρυθμίσεις που είναι πιο κατάλληλες για τις ανάγκες του.

Αποδοτικότητα: Θα πρέπει να είναι δυνατός ο ορισμός της ανομοιότητας μεταξύ των προτύπων χωρίς την ανάγκη προσπέλασης των πρωτογενών δεδομένων από τα οποία έχουν εξαχθεί τα πρότυπα. Η ανάγκη για αυτή την απαίτηση προέρχεται από το γεγονός ότι η σύνδεση με τα πρωτογενή δεδομένα μπορεί να μην είναι πάντα διαθέσιμη, λόγω π.χ., λόγων ιδιωτικότητας (όπως στο Παράδειγμα 3) ή αποδοτικότητας (όπως στο Παράδειγμα 2).

Απλότητα: Το πλαίσιο θα πρέπει να βασίζεται σε λίγες βασικές έννοιες, έτσι ώστε η εφαρμοσιμότητά του να είναι σαφής για τον τελικό χρήστη. Μία τέτοια απαίτηση εξασφαλίζει επίσης και την επεκτασιμότητα του πλαισίου.

3.2 Αναπαράσταση προτύπων

Η προσέγγισή μας στο πρόβλημα της αναπαράστασης προτύπων βασίζεται στο λογικό μοντέλο που προτάθηκε στο [70] στα πλαίσια του Ευρωπαϊκού έργου *PANDA* [65]. Ωστόσο, χρησιμοποιούμε μόνο τα τμήματα του μοντέλου που είναι σχετικά με τους σκοπούς μας.

Ακολουθώντας αυτό το μοντέλο, θεωρούμε ένα απλό σύστημα για τον ορισμό των *τύπων δεδομένων (pattern types)* - η ακριβής επιλογή των τύπων, ωστόσο, δεν επηρεάζει τον συλλογισμό μας. Το μοντέλο θεωρεί ένα σύνολο από *βασικούς τύπους (base types)*, όπως *Int*, *Real*, *Boolean* και *String*, και ένα σύνολο από *κατασκευαστές τύπων (type constructors), όπως list* ($< \ldots >$), *set* ($\{\ldots\}$), *array* ($[\ldots]$) και *tuple* ($(\ldots)$). Ας ονομάσουμε $\mathcal{T}$ το *σύνολο των τύπων* που περιέχουν όλους τους βασικούς τύπους και όλους τους τύπους που μπορούν να παραχθούν από αυτούς μέσω επαναλαμβανόμενης εφαρμογής των κατασκευαστών τύπων. Τύποι στους οποίους έχει αποδοθεί ένα (μοναδικό) όνομα ονομάζονται *ονομαστικοί τύποι (named types)*. Μερικά απλά παραδείγματα τύπων είναι:

$\{Int\}$	(σύνολο από Int)
$XYPair$ `= (x:Int,y:Int)`	(ονομαστικός τύπος πλειάδας με γνωρίσματα x και y)
$<XYPair>$	(λίστα από $XYPair$)

Ορισμός 3 (Τύπος προτύπων) *Ένας τύπος προτύπων είναι ένα ονομαστικό ζεύγος $PT = (SS, MS)$, όπου SS είναι η δομική συνιστώσα (structure schema) και MS είναι η ποσοτική συνιστώσα (measure schema). Τα SS και MS είναι τύποι στο $\mathcal{T}$. Ένας τύπος προτύπων PT ονομάζεται σύνθετος (complex) εάν η δομική του συνιστώσα SS περιέχει κάποιον άλλο τύπο προτύπων, αλλιώς ονομάζεται απλός (simple).*

Η *δομική συνιστώσα* SS ορίζει το χώρο των προτύπων περιγράφοντας τη δομή των προτύπων τα οποία είναι στιγμιότυπα του συγκεκριμένου τύπου προτύπων. Η *ποσοτική συνιστώσα* MS περιγράφει μέτρα που συσχετίζουν τα πρότυπα με τα πρωτογενή δεδομένα από τα όποία προέρχονται, δηλαδή, ποσοτικοποιεί το πόσο καλά τα πρότυπα περιγράφουν τα αντίστοιχα πρωτογενή δεδομένα.

Ένα πρότυπο (pattern) είναι ένα στιγμιότυπο ενός τύπου προτύπων, συνεπώς περιέχει τόσο τη δομική όσο και την ποσοτική συνιστώσα του τύπου προτύπων στον οποίο ανήκει. Θεωρώντας ότι κάθε βασικός τύπος B συσχετίζεται με ένα σύνολο τιμών $dom(B)$, άμεσα προκύπτει ο ορισμός τιμών για κάθε τύπο στο $\mathcal{T}$. Ανάλογα με τον τύπο του, ένα πρότυπο μπορεί να είναι είτε απλό είτε σύνθετο.

Ορισμός 4 (Πρότυπο) *Έστω $PT = (SS, MS)$ είναι ένας τύπος προτύπων. Ένα πρότυπο p, στιγμιότυπο του PT, ορίζεται ως $p = (s, m)$, όπου p είναι το αναγνωριστικό του προτύπου, s (η δομή του p, που συμβολίζεται και ως $p.s$) είναι η τιμή για τον τύπο SS, και m (το μέτρο του p, που συμβολίζεται και ως $p.m$) είναι η τιμή για τον τύπο MS.*

Για να δείξουμε την ευελιξία του μοντέλου αναπαράστασης του $PANDA$, παραθέτουμε τρία διαφορετικά σχήματα αναπαράστασης για συστάδες. Επιλέγουμε συστάδες για το παράδειγμά μας, επειδή οι διαφορετικοί αλγόριθμοι συσταδοποίησης παρέχουν μία ποικιλία περιγραφών για τις συστάδες και έτσι, από την σκοπιά της μοντελοποίησης, οι συστάδες εμφανίζουν εξαιρετικό ενδιαφέρον.

Πράγματι, υπάρχει μία πληθώρα διαφορετικών αναπαραστάσεων για τις συστάδες [30]: οι ιεραρχικοί αλγόριθμοι συσταδοποίησης περιγράφουν τις συστάδες ως σύνολα από αντικείμενα, οι αλγόριθμοι συσταδοποίησης σε μετρικό χώρο, όπως ο k-means, περιγράφουν τις συστάδες ως γεωμετρικά σχήματα, οι αλγόριθμοι που βασίζονται στην πυκνότητα όπως ο EM, περιγράφουν τις συστάδες μέσω συναρτήσεων πυκνότητας πιθανότητας κ.ο.κ.

Αυτό που περιγράφουμε στη συνέχεια δεν εξαντλεί τις πιθανές αναπαραστάσεις για τα μοντέλα συστάδων. Αντιθέτως, τα ακόλουθα παραδείγματα στοχεύουν μόνο στην ανάδειξη της λειτουργικότητας του πλαισίου $PANDA$ και δε θα πρέπει να θεωρηθούν ως τα μοναδικά ή τα πιο σωστά σχήματα αναπαράστασης συστάδων.

Παρέχουμε τρία υποψήφια σχήματα μοντελοποίησης για: (α) συστάδες που έχουν εξαχθεί μέσω κάποιου *αλγόριθμου συσταδοποίησης μετρικού χώρου* όπως ο k-means, (β) συστάδες που έχουν εξαχθεί μέσω κάποιου *αλγορίθμου συσταδοποίησης με βάση την πυκνότητα* όπως ο EM και (γ) συστάδες που έχουν εξαχθεί μέσω κάποιου *ιεραρχικού αλγορίθμου συσταδοποίησης*. Στην πρώτη περίπτωση, οι συστάδες μοντελοποιούνται ως σφαίρες (τις ονομάζουμε $EuclideanClusterType$ λόγω του μετρικού χώρου), στη δεύτερη περίπτωση οι συστάδες περιγράφονται μέσω κάποιας συνάρτησης πυκνότητας πιθανότητας (τις ονομάζουμε $DensityClusterType$), ενώ στην τρίτη περίπτωση οι συστάδες μοντελοποιούνται ως σύνολα αντικειμένων (τις ονομάζουμε $HierarchicalClusterType$). Για κάθε έναν από τους παραπάνω τύπους, περιγράφουμε επίσης έναν αντίστοιχο ενδεικτικό σύνθετο τύπο προτύπων.

Παράδειγμα 6 (*$EuclideanClusterType$* **και** *$PartitioningEuclideanClustering$*) *Μία συστάδα τύπου $EuclideanClusterType$ σε κάποιο D-διάστατο χώρο, όπως οι συστάδες που προκύπτουν από τον αλγόριθμο k-means[30], μπορεί να μοντελοποιηθεί μέσω ενός κέντρου και μίας ακτίνας, τα οποία και αποτελούν τη* δομική συνιστώσα *της συστάδας. Ως* ποσοτική συνιστώσα *θα μπορούσαμε να θεωρήσουμε την υποστήριξη της συστάδας, δηλαδή, το ποσοστό των αντικειμένων που ανήκουν στη συστάδα, και τη μέση απόσταση μεταξύ των αντικειμένων-μελών της συστάδας. Δηλαδή:*

$$EuclideanCluster =$$

$$(SS : (center : [Real]_1^D, radius : Real),$$

$$MS : (supp : Real, avgdist : Real))$$

Αν υποθέσουμε $D = 3$, ένα πιθανό στιγμιότυπο αυτού του τύπου θα μπορούσε να είναι ως ακολούθως:

$$p407 = (s : (center = [0.75, 1.25, 0.46], radius = 0.24), (m : supp = 0.13, avgdist = 0.17))$$

Ένα $PartitioningEuclideanClustering$ πρότυπο μπορεί να ορισθεί ως σύνθεση προτύπων τύπου $EuclideanCluster$. Ειδικά στην περίπτωση ενός αυστηρού αλγορίθμου διαμεριστικής συσταδοποίησης (hard partitioning clustering) όπως ο

k-means, ένα EuclideanClustering πρότυπο μπορεί απλά να μοντελοποιηθεί ως ένα σύνολο από πρότυπα EuclideanCluster χωρίς κάποια ποσοτική συνιστώσα:

$$PartitioningEuclideanClustering = (SS : \{EuclideanCluster\}, MS : \perp)$$

□

Παράδειγμα 7 (*DensityClusterType* **και** *DensityBasedClustering*) *Ένας αλγόριθμος συσταδοποίησης βασισμένος στην πυκνότητα, όπως ο EM [30], παράγει συστάδες που μπορούν να περιγραφούν μέσω μίας π.χ., Gaussian κατανομής πιθανότητας. Συνεπώς μπορούν να μοντελοποιηθούν μέσω ενός μέσου (mean) και μίας τυπικής απόκλισης (standard deviation) σε κάθε διάσταση - αυτά αποτελούν και τη δομική συνιστώσα της συστάδας. Ως ποσοτική συνιστώσα θα μπορούσαμε να θεωρήσουμε την υποστήριξη της συστάδας (cluster support), δηλαδή, το ποσοστό του πληθυσμού που καλύπτεται από την εν λόγω συστάδα. Σύμφωνα με τα παραπάνω, λοιπόν:*

$$DensityBasedCluster =$$

$$SS : (mean : [Real]_1^D, stdDev : [Real]_1^D),$$

$$MS : (supp : Real))$$

Θεωρώντας $D = 2$, ένα πιθανό στιγμιότυπο αυτού του τύπου θα μπορούσε να είναι ως εξής:

$$p111 = (s : (mean = [15.5, 41.4], stdDev = [3.6, 4.7]), m : (supp = 0.33))$$

Ένα πρότυπο τύπου DensityBasedClustering μπορεί να μοντελοποιηθεί ως ένα σύνολο από πρότυπα τύπου DensityBasedCluster χωρίς κάποια ποσοτική συνιστώσα:

$$PartitioningDensityBasedClustering = (SS : \{DensityBasedCluster\}, MS : \perp)$$

□

Παράδειγμα 8 (*HierarchicalClusterType* **και** *HierarchicalClustering*) *Μία συστάδα που προέρχεται από κάποιο ιεραρχικό αλγόριθμο περιγράφεται συνήθως με βάση το σύνολο αντικειμένων που την αποτελούν (δομική συνιστώσα). Ως μετρική συνιστώσα θα μπορούσε να θεωρηθεί το πλήθος των στοιχείων του συνόλου αυτού. Συνεπώς:*

$$Point = (SS : (coords : [Real]_1^D), MS : \perp)$$

$$HierarchicalCluster = (SS : \{Point\}, MS : (supp : Real))$$

Ένα πρότυπο τύπου HierarchicalClustering μπορεί να μοντελοποιηθεί ως ένα σύνολο από πρότυπα τύπου HierarchicalCluster χωρίς κάποια ποσοτική συνιστώσα:

$$HierarchicalClustering = (SS : \{HierarchicalCluster\}, MS : \perp$$

□

Να σημειώσουμε και πάλι πως η διάκριση μεταξύ απλών και σύνθετων προτύπων βασίζεται στο κατά πόσο η δομική τους συνιστώσα ορίζεται πάνω σε πρωτογενή δεδομένα ή πάνω σε άλλα πρότυπα, αντίστοιχα. Στη δεύτερη περίπτωση, τα πρότυπα είναι σύνθετα, π.χ., μία συστάδα από κανόνες συσχέτισης. Η απλούστερη μορφή σύνθετων προτύπων περιέχει μία ιεραρχία 2 επιπέδων, δηλαδή, το σύνθετο πρότυπο και τα επιμέρους απλά πρότυπα που το αποτελούν. Ο ορισμός ενός σύνθετου προτύπου ωστόσο, επιτρέπει την πολλαπλή εμφώλευση μέσα στη δομική συνιστώσα και έτσι, ιεραρχίες οποιουδήποτε επιπέδου υποστηρίζονται. Ένα ενδεικτικό παράδειγμα είναι αυτό ενός δικτυακού τόπου: τυπικά, τα περιεχόμενα ενός δικτυακού τόπου είναι οργανωμένα σε κατηγορίες, κάθε κατηγορία περιέχει ένα σύνολο από ιστοσελίδες και κάθε ιστοσελίδα περιγράφεται από ένα σύνολο λέξεων - κλειδιά. Το παράδειγμα αυτό μπορεί να ενταχθεί στην λογική των απλών - σύνθετων προτύπων του $PANDA$ ως εξής: Ένας *δικτυακός τόπος* αναπαριστά ένα σύνθετο πρότυπο που αποτελείται από άλλα πρότυπα, τις κατηγορίες. Μία *κατηγορία* αντιστοιχεί επίσης σε ένα σύνθετο πρότυπο που αποτελείται από άλλα πρότυπα, τις ιστοσελίδες. Μία *ιστοσελίδα* αναπαριστά με την σειρά της ένα άλλο σύνθετο πρότυπο το οποίο αποτελείται από άλλα πρότυπα, τις λέξεις κλειδιά. Μία *λέξη - κλειδί*, τέλος, αντιστοιχεί σε ένα απλό πρότυπο, και μπορεί να περιγραφεί μέσω κάποιου ονόματος (π.χ., " υπολογιστής ") και κάποιου βάρους (π.χ., 0.5) που ποσοτικοποιεί την σημαντικότητα της λέξης στην σελίδα.

Από τη στιγμή που κάποιο πρόβλημα σύγκρισης προτύπων μπορεί να ενταχθεί στη λογική των απλών/σύνθετων προτύπων, μπορεί να αντιμετωπιστεί κατευθείαν μέσω του πλαισίου $PANDA$.

Ο Πίνακας 3.1 συνοψίζει τα σύμβολα που χρησιμοποιούνται σε αυτό το κεφάλαιο.

Σύμβολο	**Περιγραφή**
p	απλό πρότυπο
cp	σύνθετο πρότυπο
dis_{struct}	συνάρτηση δομικής ανομοιότητας
dis_{meas}	συνάρτηση ποσοτικής ανομοιότητας
Συνδυαστής (Comb)	συνάρτηση συνδυασμού των δομικών και ποσοτικών ανομοιοτήτων
Τύπος ταιριάσματος (Matcher)	καθορίζει πως θα ταιριάξουν τα επιμέρους πρότυπα των σύνθετων προτύπων
Λογική συνάθροισης (Aggr)	καθορίζει πως θα συναθροιστούν τα σκορ των ταιριασμένων επιμέρους προτύπων

Πίνακας 3.1: Λίστα συμβόλων για το Κεφάλαιο 3

3.3 Το πλαίσιο $PANDA$

Σε αυτή την ενότητα, περιγράφουμε το πλαίσιο $PANDA$ για την αποτίμηση της ανομοιότητας μεταξύ δύο προτύπων p_1, p_2 του ιδίου τύπου προτύπων PT.

Οι βασικές αρχές στις οποίες βασίζεται το πλαίσιο μας είναι οι ακόλουθες:

1. Η ανομοιότητα μεταξύ δύο προτύπων θα πρέπει να δίνει μία *τιμή*, κανονικοποιημένη στο διάστημα [0..1] (μεγαλύτερη τιμή αντιστοιχεί σε μεγαλύτερη ανομοιότητα).

2. Η ανομοιότητα μεταξύ δύο σύνθετων προτύπων θα πρέπει να εξαρτάται (αναδρομικά) από την ανομοιότητα των επιμέρους προτύπων που τα αποτελούν.

3. Η ανομοιότητα μεταξύ δύο προτύπων θα πρέπει να υπολογίζεται με βάση τόσο την ανομοιότητα των δομικών τους συνιστωσών όσο και την ανομοιότητα των ποσοτικών τους συνιστωσών.

Η πρώτη αρχή, η κανονικοποίηση των τιμών της ανομοιότητας, προσφέρει μία καλύτερη και πιο διαισθητική ερμηνεία των αποτελεσμάτων.

Η δεύτερη αρχή παρέχει την όλη ευελιξία του πλαισίου μας. Σημειώστε πως στην περίπτωση των σύνθετων προτύπων, θα μπορούσε κανείς να χρησιμοποιήσει αυθαίρετα μοντέλα για τη σύγκριση. Ωστόσο, είναι χρήσιμο και, την ίδια στιγμή, επαρκές για πρακτικούς λόγους, να θεωρήσουμε λύσεις που αποσυνθέτουν το δύσκολο πρόβλημα της σύγκρισης σύνθετων προτύπων σε πιο εύκολα υπό-προβλήματα όπως αυτό της σύγκρισης απλών προτύπων, και στη συνέχεια να συναθροίσουμε έξυπνα τις τιμές των επιμέρους λύσεων σε μία συνολική τιμή.

Η τρίτη αρχή είναι μία άμεση συνέπεια της προσέγγισής μας που επιτρέπει αυθαίρετα σύνθετες δομές προτύπων. Δεδομένου ότι η δομή ενός σύνθετου προτύπου μπορεί να περιέχει ποσοτικές συνιστώσες από τα επιμέρους πρότυπα, αγνοώντας τη δομική ανομοιότητα μπορούν εύκολα να προκύψουν λανθασμένα αποτελέσματα, όπως π.χ., να συγκρίνουμε δύο σημασιολογικά αντικρουόμενα πρότυπα. Για να αντιμετωπίσουμε αυτές τις περιπτώσεις, εισάγουμε την έννοια της συνάρτησης συνδυασμού (combining function που συνδυάζει την τιμή της δομικής και της ποσοτικής ανομοιότητας μόνο εφόσον υπάρχει κάποια δομική συμβατότητα μεταξύ των προτύπων. Ένα άλλο κίνητρο που διέπει την αρχή αυτή προκύπτει από την ανάγκη κατασκευής ενός *αποδοτικού* πλαισίου, το οποίο δεν θα απαιτεί την προσπέλαση των πρωτογενών συνόλων δεδομένων για την εκτίμηση της ανομοιότητας, μέσω π.χ., των κοινών τους στιγμιότυπων. Για το σκοπό αυτό, αξιοποιούμε όλα τα κομμάτια της πληροφορίας που είναι διαθέσιμα στον χώρο των προτύπων, δηλαδή τη δομική περιγραφή του χώρου των προτύπων και τα ποσοτικά μέτρα που συσχετίζουν το χώρο των προτύπων με το χώρο των πρωτογενών δεδομένων, από τα οποία έγινε η εξαγωγή των προτύπων.

Επιπλέον, για κάθε τύπο προτύπων που μας ενδιαφέρει θα πρέπει να οριστεί ένας τουλάχιστον τελεστής ανομοιότητας. Την ίδια στιγμή, ωστόσο, θα πρέπει να μπορούμε να ορίζουμε πολλαπλούς τελεστές ανομοιότητας για τον ίδιο τύπο προτύπων.

Στις επόμενες υποενότητες, πρώτα περιγράφουμε πώς οι παραπάνω αρχές εφαρμόζονται στην (απλή) περίπτωση της σύγκρισης απλών προτύπων (Ενότητα 3.3.1) και στη συνέχεια, δείχνουμε πώς μπορούν να γενικευτούν για να καλύψουν την (γενική) περίπτωση της σύγκρισης μεταξύ σύνθετων προτύπων (Ενότητα 3.3.2).

3.3.1 Ανομοιότητα μεταξύ απλών προτύπων

Η ανομοιότητα μεταξύ δύο απλών προτύπων p_1, p_2 του ιδίου απλού τύπου προτύπων PT βασίζεται σε τρία βασικά συστατικά:

- μία συνάρτηση *δομικής ανομοιότητας*, dis_{struct}, η οποία αποτιμά την ανομοιότητα των αντίστοιχων δομικών συνιστωσών $p_1.s$ και $p_2.s$,

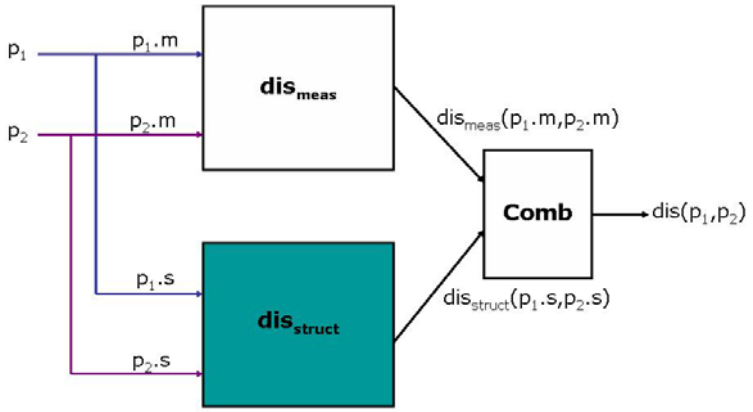
- μία συνάρτηση *ποσοτικής ανομοιότητας*, dis_{meas}, η οποία αποτιμά την ανομοιότητα των αντίστοιχων ποσοτικών συνιστωσών $p_1.m$ και $p_2.m$, και

- μία συνάρτηση *συνδυασμού*, $Comb$, που συνδυάζει τις τιμές της δομικής και της ποσοτικής ανομοιότητας σε μία συνολική τιμή που αντανακλά τη συνολική ανομοιότητα μεταξύ των συγκρινόμενων (απλών) προτύπων.

Συμπερασματικά, η ανομοιότητα μεταξύ δύο προτύπων ορίζεται ως εξής:

$$dis(p_1, p_2) = Comb(dis_{struct}(p_1.s, p_2.s), dis_{meas}(p_1.m, p_2.m)) \tag{3.1}$$

Η γενική ιδέα της όλης διαδικασίας αποτίμησης της ανομοιότητας παρουσιάζεται στο Σχήμα 3.1.



Σχήμα 3.1: Αποτίμηση της ανομοιότητας μεταξύ (απλών) προτύπων

Εάν τα πρότυπα p_1, p_2 έχουν ίδια δομή, τότε $dis_{struct}(p_1.s, p_2.s) = 0$ και η ανομοιότητά τους εξαρτάται αποκλειστικά από την ανομοιότητα των ποσοτικών τους συνιστωσών. Στη γενική περίπτωση, ωστόσο, τα πρότυπα που συγκρίνονται μπορεί να έχουν διαφορετική δομή και τότε υπάρχουν δύο εναλλακτικές λύσεις :

- Οι δομικές συνιστώσες είναι σε κάποιο βαθμό συμβατές και η τιμή της δομικής ανομοιότητας, $dis_{struct}(p_1.s, p_2.s)$, αντανακλά το επιπλέον κόστος ανομοιότητας που απαιτείται σε σχέση με την περίπτωση της ίδιας δομής.

- Οι δομικές συνιστώσες είναι τελείως ασύμβατες (με κάποια έννοια που εξαρτάται από την εκάστοτε εφαρμογή), δηλαδή, $dis_{struct}(p_1.s, p_2.s) = 1$. Σε αυτή την περίπτωση, αδιαφορώντας για την τιμή της ποσοτικής ανομοιότητας, απαιτούμε επιπλέον ότι η συνολική τιμή της ανομοιότητας θα πρέπει να είναι η μέγιστη, δηλαδή, $dis(p_1, p_2) = 1$. Αυτός ο περιορισμός επιβάλλεται για να αντιμετωπιστούν περιπτώσεις όπου δύο τελείως διαφορετικά πρότυπα μπορεί να θεωρηθούν παρόμοια λόγω μικρής διαφοράς στις ποσοτικές τους συνιστώσες.

Να τονίσουμε πως στο $PANDA$ οι δομικές συνιστώσες παίζουν τον πρωταρχικό ρόλο στη διαδικασία της αποτίμησης της ανομοιότητας. Αυτό είναι λογικό, καθώς προκειμένου να συγκρίνουμε τις ποσοτικές συνιστώσες δύο προτύπων θα πρέπει να υπάρχει κάποιο είδος συμβατότητας μεταξύ των δομών τους, δηλαδή, τα πρότυπα που πρόκειται να συγκριθούν θα πρέπει να είναι "κάπως σχετικά".

Παρακάτω παρουσιάζουμε δύο παραδείγματα σύγκρισης απλών προτύπων: σύγκριση μεταξύ συχνών στοιχειοσυνόλων (frequent itemsets) που σχετίζεται με το Παράδειγμα 2 της Ενότητας 3.1 και σύγκριση μεταξύ λέξεων - κλειδιά που σχετίζεται με το Παράδειγμα 4 της Ενότητας 3.1. Αρχικά, περιγράφουμε τη δομική και ποσοτική συνιστώσα κάθε τύπου προτύπων και στη συνέχεια, παρουσιάζουμε πως αρχικοποιούνται τα διάφορα τμήματα του $PANDA$ που εμπλέκονται στη διαδικασία υπολογισμού της ανομοιότητας μεταξύ απλών προτύπων, δηλαδή τα dis_{struct}, dis_{meas} και $Comb$.

Παράδειγμα 9 (Στοιχειοσύνολα - Itemsets) *Ας θεωρήσουμε ένα σύνολο δεδομένων D που περιγράφεται από τη σχέση $R = \{I_1, I_2, \ldots, I_m\}$. Ένα στοιχειοσύνολο (itemset) IS που εξάγεται από το D μπορεί να μοντελοποιηθεί ως ένα απλό πρότυπο. Η δομή του αποτελείται από ένα σύνολο στοιχείων του R, δηλαδή, $IS.s = I_1, \ldots, I_m$, ενώ το μέτρο του είναι η υποστήριξη του στοιχειοσυνόλου, δηλαδή, το ποσοστό των πλειάδων στο D που περιέχει όλα τα στοιχεία του $IS.s$ ($IS.m = (supp : Real)$).*

Ας θεωρήσουμε τώρα το πρόβλημα της σύγκρισης των ακόλουθων δύο στοιχειοσυνόλων IS_1, IS_2 που εξάγονται μέσω των παραπάνω ρυθμίσεων:

$$IS_1 = (\{bread,\ honey,\ milk\},\ 0.1) \text{ και } IS_2 = (\{butter,\ milk\},\ 0.2)$$

- ***Δομική ανομοιότητα:*** *Για τη συνάρτηση της δομικής ανομοιότητας, θα μπορούσαμε να χρησιμοποιήσουμε την επικάλυψη των δομών τους, δηλαδή, πόσα στοιχεία είναι κοινά στα δύο στοιχειοσύνολα:*

$$dis_{struct} = 1 - \frac{IS_1.s \cap IS_2.s}{IS_1.s \cup IS_2.s} \tag{3.2}$$

Στο παράδειγμά μας, αυτό ισούται με $1 - \frac{1}{4} = 0.75$.

- ***Ποσοτική ανομοιότητα:*** *Για τη συνάρτηση της ποσοτικής ανομοιότητας, θα μπορούσαμε να θεωρήσουμε την απόλυτη διαφορά των μέτρων τους (ποσοτικών συνιστωσών τους):*

$$dis_{meas}(IS_1.m, IS_2.m) = |IS_1.m.supp - IS_2.m.supp| \tag{3.3}$$

Στο παράδειγμά μας, αυτό ισούται με $|0.1 - 0.2| = 0.1$.

- ***Συνδυαστής:*** *Τέλος, για τη συνάρτηση συνδυασμού θα μπορούσαμε να θεωρήσουμε τη μέση τιμή της δομικής και ποσοτικής ανομοιότητας, εφόσον βέβαια οι δομές των στοιχειοσυνόλων είναι σχετικές, δηλαδή,*

$$Comb(dis_{struct}, dis_{meas}) = \begin{cases} 1 & , \text{ιφ } dis_{struct} = 1 \\ \text{avg}(dis_{struct}, dis_{meas}) & , \text{οτηερωισε} \end{cases} \tag{3.4}$$

Στο παράδειγμά μας, αυτό ισούται με $(0.75 + 0.1)/2 = 0.425$.

Εάν, σε κάποιο εναλλακτικό σενάριο, ο τελικός χρήστης θεωρεί ότι η δομική ανομοιότητα είναι περισσότερο σημαντική από την ποσοτική ανομοιότητα, π.χ., με μία αναλογία 3:1, η συνάρτηση συνδυαστής θα έχει την ακόλουθη μορφή:

$$Comb(dis_{struct}, dis_{meas}) = \begin{cases} 1 & , \text{ιφ } dis_{struct} = 1 \\ \frac{3}{4} * dis_{struct} + \frac{1}{4} * dis_{meas} & , \text{οτηερωισε} \end{cases} \tag{3.5}$$

Στο παράδειγμά μας, αυτό ισούται με $\frac{3}{4} \cdot 0.75 + \frac{1}{4} \cdot 0.1 = 0.5875$.

□

Παράδειγμα 10 (Λέξεις κλειδιά - keywords) *Μία λέξη-κλειδί που εξάγεται από κάποιο κείμενο μπορεί να αναπαρασταθεί ως ένα ζεύγος* $(s = t, m = w)$*, όπου* t *είναι η λέξη (δομική συνιστώσα) και* $w \in (0, 1]$ *είναι το κανονικοποιημένο της βάρος στο κείμενο (ποσοτική συνιστώσα). Ας θεωρήσουμε δύο λέξεις - κλειδιά:* $k_1 = (s = t_1, m = w_1)$ *και* $k_2 = (s = t_2, m = w_2)$*. Όπως και με το προηγούμενο παράδειγμα, για να ορίσουμε την ανομοιότητα μεταξύ δύο λέξεων, θα πρέπει να ορίσουμε πώς οι δομικές και οι ποσοτικές τους συνιστώσες συγκρίνονται και πώς συνδυάζονται οι προκύπτουσες τιμές.*

- ***Δομική ανομοιότητα:*** *Εαν οι δύο λέξεις είναι ίδιες, τότε* $dis_{struct}(k_1, k_2) = 0$*. Όταν* $t_1 \neq t_2$*, δύο εναλλακτικές λύσεις μπορούν να επιλεγούν: αν είναι διαθέσιμη κάποια πληροφορία για τη σημασιολογία των λέξεων, τότε η* $dis_{struct}(k_1, k_2)$ *θα μπορούσε να αντιστοιχεί στη σημασιολογική απόσταση μεταξύ των* t_1 *και* t_2*, π.χ., θεωρώντας τον κοντινότερο κοινό τους πρόγονο (LCA) [10]. Εάν, από την άλλη, δεν είναι διαθέσιμη καμία τέτοια πληροφορία, τότε* $dis_{struct}(k_1, k_2) = 1$*.*

- ***Ποσοτική ανομοιότητα:*** *Μία πιθανή επιλογή για τη συνάρτηση ποσοτικής ανομοιότητας είναι η απόλυτη διαφορά των μέτρων τους, δηλαδή,* $dis_{meas}(w_1, w_2) = |w_1 - w_2|$*.*

- ***Συνδυαστής:*** *Τέλος, μία κατάλληλη συνάρτηση συνδυασμού για το παράδειγμα αυτό θα μπορούσε να είναι η ακόλουθη:*

$$\begin{aligned} dis(k_1, k_2) &= dis_{struct}(k_1.s, k_2.s) + dis_{meas}(k_1.m, k_2.m) \\ &\quad - dis_{struct}(k_1.s, k_2.s) * dis_{meas}(k_1.m, k_2.m) \\ &= dis_{struct}(t_1, t_2) + dis_{meas}(w_1, w_2) - dis_{struct}(t_1, t_2) * dis_{meas}(w_1, w_2) \end{aligned} \tag{3.6}$$

από την οποία προκύπτει ότι $dis(k_1, k_2) = 1$ *όπου* $dis_{struct}(t_1, t_2) = 1$*, και* $dis(k_1, k_2) = dis_{meas}(w_1, w_2)$ *όταν* $dis_{struct}(t_1, t_2) = 0$*.*

□

Τα παραπάνω παραδείγματα δείχνουν την ευελιξία του πλαισίου $PANDA$: τόσο η δομική όσο και η ποσοτική συνάρτηση ανομοιότητας, όπως και η συνάρτηση συνδυασμού μπορούν εύκολα να προσαρμοστούν στις συγκεκριμένες απαιτήσεις του χρήστη/εφαρμογής. Επειδή ακριβώς υπάρχουν πολλοί τρόποι αρχικοποίησης αυτών των συναρτήσεων, διάφορες συναρτήσεις ανομοιότητας για τη σύγκριση (απλών) προτύπων προκύπτουν στα πλαίσια του πλαισίου $PANDA$, διευκολύνοντας τον τελικό χρήστη να επιλέξει την πιο κατάλληλη συνάρτηση ανάλογα με τις ανάγκες του.

3.3.2 Ανομοιότητα μεταξύ σύνθετων προτύπων

Παρότι για λόγους αρχής, κάποιος θα μπορούσε να θεωρήσει απλά πρότυπα με αυθαίρετα σύνθετες δομικές συνιστώσες, αυτό θα ανάγκαζε τις συναρτήσεις ανομοιότητας να είναι σύνθετες και μη επαναχρησιμοποιούμενες. Ανάμεσα στις απαιτήσεις

που δηλώθηκαν στην Ενότητα 3.1, αυτή η μονολιθική προσέγγιση θα συμμορφωνόταν μόνο με την απαίτηση για αποδοτικότητα, χωρίς να καλύπτει τις υπόλοιπες απαιτήσεις. Στο *PANDA* αναζητούμε μία αρθρωτή προσέγγιση η οποία, εξ ορισμού, εγγυάται ευελιξία, απλότητα και επαναχρησιμοποίηση. Επιπλέον, όπως θα αναφερθεί στην συνέχεια, αυτή η προσέγγιση δεν αποκλείει την πιθανότητα αποδοτικών υλοποιήσεων.

Ακολουθώντας το μοντέλο προτύπων που περιγράφηκε στην Ενότητα 3.2, η ανομοιότητα μεταξύ σύνθετων προτύπων ορίζεται αναδρομικά με βάση την ανομοιότητα των επιμέρους προτύπων που τα αποτελούν. Η δομή των σύνθετων προτύπων παίζει εδώ κύριο ρόλο, καθώς η διαδικασία σύγκρισης βασίζεται στις δομικές συνιστώσες των προς σύγκριση προτύπων.

Χωρίς απώλεια της γενικότητας, σε ότι ακολουθεί θεωρούμε ότι τα επιμέρους πρότυπα, $p^1, p^2, \ldots, p^N$, ενός σύνθετου προτύπου cp περιγράφουν πλήρως τη δομή του cp (καμία επιπλέον πληροφορία δεν υπάρχει στο $cp.s$) και ότι ορίζουν ένα σύνολο ($cp.s = \{p^1, p^2, \ldots, p^N\}$).

Η *δομική ανομοιότητα* μεταξύ δύο σύνθετων προτύπων ($cp_1 = \{p_1^1, p_1^2, \ldots, p_1^{N_1}\}$) και ($cp_2 = \{p_2^1, p_2^2, \ldots, p_2^{N_2}\}$) μπορεί εύκολα να προσαρμοστεί σε συγκεκριμένες ανάγκες/περιορισμούς καθώς στηρίζεται σε δύο θεμελιώδεις έννοιες:

- τον *τύπο ταιριάσματος (matching type)*, ο οποίος καθορίζει τον τρόπο ταιριάσματος των επιμέρους προτύπων των cp_1, cp_2, και
- τη *λογική συνάθροισης (aggregation logic)*, η οποία καθορίζει πως θα συναθροιστούν τα σκορ ανομοιότητας των ταιριασμένων επιμέρους προτύπων σε ένα σκορ που αναπαριστά την συνολική τιμή της ανομοιότητας μεταξύ των δομών των σύνθετων προτύπων.

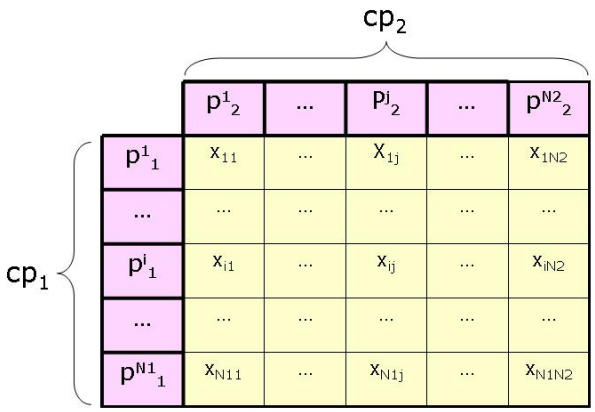
3.3.2.1 Τύπος ταιριάσματος

Όπως έχουμε ήδη αναφέρει, ένα σύνθετο πρότυπο μπορεί τελικά να αποσυντεθεί σε ένα αριθμό από επιμέρους πρότυπα. Έτσι, όταν συγκρίνουμε δύο σύνθετα πρότυπα cp_1, cp_2 qreiaz'omaste 'enan tr'opo gia na susqet'isoume/tairi'axoume ta epim'erous pr'otupa. Gia to skop'o aut'o, eis'agoume thn 'ennoia tou *t'upou tairi'asmatos*, o opo'ios or'izei p'ws ta epim'erous pr'otupa tou cp_1 (cp_2) ταιριάζονται με τα επιμέρους πρότυπα του cp_2 (cp_1, αντίστοιχα), λαμβάνοντας υπόψη συγκεκριμένες απαιτήσεις του χρήστη/εφαρμογής.

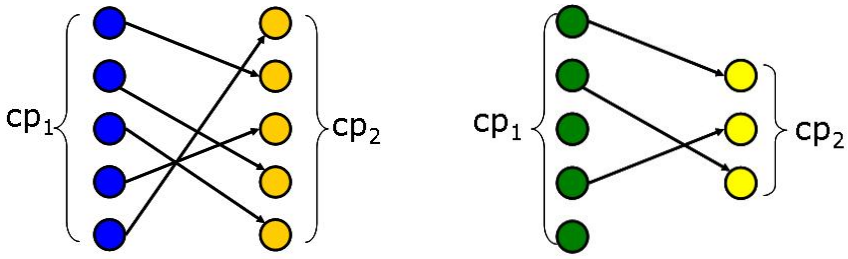
Ένα ταίριασμα μεταξύ των σύνθετων προτύπων cp_1, cp_2 μπορεί να αναπαρασταθεί μέσω ενός *πίνακα ταιριασμάτων* $\mathbf{X}_{N_1 \times N_2} = (x_{ij})$, όπου κάθε στοιχείο $x_{ij} \in [0, 1]$ ($i = 1, \ldots, N_1$, $j = 1, \ldots, N_2$) αναπαριστά το ποσό του ταιριάσματος του i-οστου επιμέρους προτύπου του cp_1, p_1^i, με το j-οστό επιμέρους προτύπου του cp_2, p_2^j. Ένας τέτοιος πίνακας παρουσιάζεται στο Σχήμα 3.2.

Στην ουσία, ένας *τύπος ταιριάσματος* είναι ένα σύνολο από περιορισμούς στους συντελεστές x_{ij} έτσι ώστε μόνο κάποια από τα δυνατά ταιριάσματα να είναι έγκυρα. Παρακάτω, περιγράφουμε μερικά παραδείγματα τύπων ταιριασμάτων:

- **1–1 ταίριασμα:** Σε αυτή την περίπτωση, κάθε επιμέρους πρότυπο του cp_1 μπορεί να ταιριαστεί με ένα το πολύ επιμέρους πρότυπο του cp_2 και το αντίθετο. Αν $N_1 = N_2$, υπάρχει ένα πλήρες ταίριασμα μεταξύ των cp_1 και cp_2 όπως στο Σχήμα 3.3 (α). Ένα μερικό ταίριασμα μπορεί να συμβεί αν $N_1 \neq N_2$ όπως στο Σχήμα 3.3 (β).



Σχήμα 3.2: Ο πίνακας ταιριασμάτων μεταξύ των σύνθετων προτύπων cp_1, cp_2



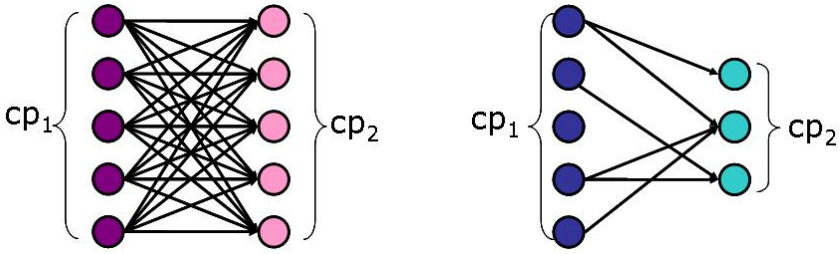
(α) Πλήρες ταίριασμα (β) Μερικό ταίριασμα

Σχήμα 3.3: 1–1 ταίριασμα

Το 1–1 ταίριασμα αντιστοιχεί στο ακόλουθο σύνολο από περιορισμούς:

$$\begin{aligned} x_{ij} &\in \{0,1\} \\ \sum_{i=1}^{N_1} x_{ij} &\leq 1, \quad \forall j \\ \sum_{j=1}^{N_2} x_{ij} &\leq 1, \quad \forall i \\ \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} x_{ij} &= \min\{N_1, N_2\}, \quad \forall i,j \end{aligned} \tag{3.7}$$

- **N–M ταίριασμα:** Σε αυτή την περίπτωση, κάθε επιμέρους πρότυπο του cp_1 μπορεί να ταιριαστεί με παραπάνω από ένα επιμέρους πρότυπα του cp_2, και το αντίθετο. Στην ακραία περίπτωση, κάθε επιμέρους πρότυπο του cp_1 ταιριάζεται με όλα τα επιμέρους πρότυπα του cp_2, όπως στο Σχήμα 3.4 (α). Στη γενική περίπτωση, ωστόσο, ένα μερικό ταίριασμα μπορεί να υπάρξει, όπως στο Σχήμα 3.4 (β).



(α) Πλήρες ταίριασμα (β) Μερικό ταίριασμα

Σχήμα 3.4: N–M ταίριασμα

Το N–M ταίριασμα αντιστοιχεί στο ακόλουθο σύνολο από περιορισμούς:

$$\begin{aligned} x_{ij} &\in \{0,1\} \\ \sum_{i=1}^{N_1} x_{ij} &\leq N_2, \quad \forall j \\ \sum_{j=1}^{N_2} x_{ij} &\leq N_1, \quad \forall i \\ \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} x_{ij} &\leq N_1 * N_2, \quad \forall i,j \end{aligned} \tag{3.8}$$

- **EMD ταίριασμα:** Αυτός ο τυπος ταιριάσματος, που εισήχθηκε για τον ορισμό του Earth Movers Distance (EMD) [71, 42], διαφέρει από τους προηγούμενους τύπους ταιριάσματος καθώς κάθε επιμέρους πρότυπο p συσχετίζεται επιπλέον και με κάποιο βάρος w.

  Συνεπώς, τα σύνθετα πρότυπα cp_1, cp_2 παίρνουν τώρα τη μορφή:

  $$cp_1 = \{(p_1^1, w_1^1), (p_1^2, w_1^2), \ldots, (p_1^{N_1}, w_1^{N_1})\}$$

  $$cp_2 = \{(p_2^1, w_2^1), (p_2^2, w_2^2), \ldots, (p_2^{N_2}, w_2^{N_2})\}$$

  και το EMD ορίζεται ως εξής:

  $$W(cp_1, cp_2, X) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} x_{ij} d_{ij}$$

  όπου d_{ij} είναι κάποιο μέτρο ανομοιότητας μεταξύ των p_1^i και p_2^j.

Το EMD ταίριασμα αντιστοιχεί στο ακόλουθο σύνολο περιορισμών:

$$\begin{aligned} x_{ij} &\geq 0 \\ \sum_{i=1}^{N_1} x_{ij} &\leq w_2^j, \quad \forall j \\ \sum_{j=1}^{N_2} x_{ij} &\leq w_1^i, \quad \forall i \\ \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} x_{ij} &= \min(\sum_{i=1}^{N_1} w_1^i, \sum_{j=1}^{N_2} w_2^j) \end{aligned} \tag{3.9}$$

- **DTW ταίριασμα:** Το Dynamic Time Warping (DTW) είναι μία ευρέως χρησιμοποιούμενη τεχνική για εκτίμηση της ομοιότητας χρονοσειρών [18]. Η ιδιαιτερότητά της έγκειται στο γεγονός ότι βρίσκει το καλύτερο ταίριασμα μεταξύ δύο χρονοσειρών X, Y επιτρέποντας τοπικές παραμορφώσεις, συγκεκριμένα εκτάσεις (stretch) ή συρρικνώσεις (shrink) στον άξονα του χρόνου. Μία τέτοια ευθυγράμμιση ονομάζεται warping path. Η διαδικασία της εύρεσης των καλύτερων ευθυγραμμίσεων περιλαμβάνει την εύρεση όλων των δυνατών warping paths μεταξύ της X και της Y και την επιλογή εκείνου που ελαχιστοποιεί την ολική απόσταση. Το DTW δίνεται από την εξίσωση:

$$DTW(X,Y) = \min_{\forall X',Y' \ s.t. \ |X'|=|Y'|} L_1(X',Y') \tag{3.10}$$

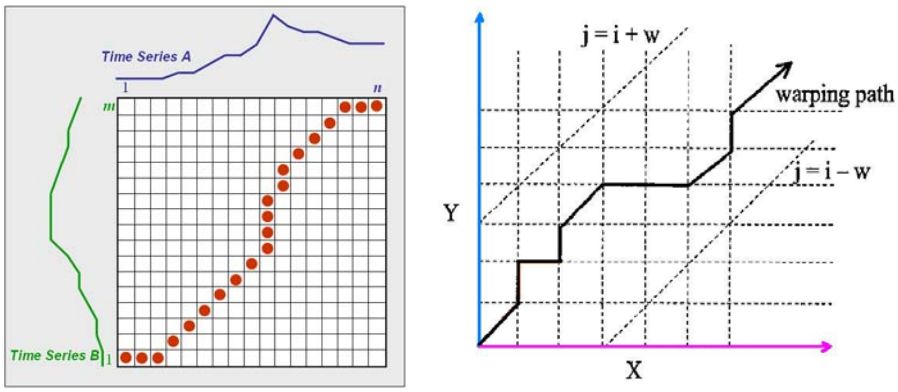
  όπου X', Y' είναι οι νέες ακολουθίες που προκύπτουν από τις αρχικές ακολουθίες X, Y με την επανάληψη των στοιχείων τους. Ένα αντιπροσωπευτικό παράδειγμα φαίνεται στο Σχήμα 3.5 (α).

  Μερικές φορές, για λόγους αποδοτικότητας, ορίζεται ένα φραγμένο παράθυρο warping μεγέθους w, το οποίο ελέγχει την απόσταση μεταξύ ενός χρονικού σημείου της X και του ταιριάσματός του στην Y. Σε αυτή την περίπτωση, για κάθε συντελεστή $x_{i,j}$ στο μονοπάτι warping, θα πρέπει να ισχύει: $|i - j| \leq w$. Ένα τέτοιο παράδειγμα φαίνεται στο Σχήμα 3.5 (β).

Πριν κλείσουμε την ενότητα αυτή, θα πρέπει να αναφέρουμε πως οι συναρτήσεις ανομοιότητας, είτε άμεσα είτε έμμεσα, βασίζονται σε κάποιο είδος ταιριάσματος: Για παράδειγμα, το 1–1 ταίριασμα αποτελεί βασικό στοιχείο στη σύγκριση γράφων. Επίσης, κατά τη σύγκριση κατανομών ή ιστογραμμάτων ένα 1–1 ταίριασμα εφαρμόζεται, π.χ., η διαφορά Kullback-Leibler (KL divergence) μετρά την διαφορά μεταξύ δύο κατανομών συγκρίνοντας, για κάθε bucket, την πιθανότητα στην πρώτη και τη δεύτερη κατανομή [35]. Ένα άλλο παράδειγμα είναι η τεχνική της ιεραρχικής συσταδοποίησης. Στον αλγόριθμο πλήρους διασύνδεσης (complete linkage), εφαρμόζεται ένα είδος N–M ταιριάσματος μεταξύ των μελών των συστάδων έτσι ώστε να αποφασιστεί ποιες συστάδες θα ενωθούν ή θα διασπαστούν στο επόμενο βήμα του αλγορίθμου [30].

3.3.2.2 Λογική συνάθροισης

Για τον υπολογισμό της ολικής ανομοιότητας μεταξύ δύο σύνθετων προτύπων, οι τιμές της ανομοιότητας των ταιριασμένων επιμέρους προτύπων θα πρέπει να



(α) Ελεύθερο (μη-φραγμένο) DTW (β) Φραγμένο DTW

Σχήμα 3.5: Ταίριασμα Dynamic Time Warping (DTW)

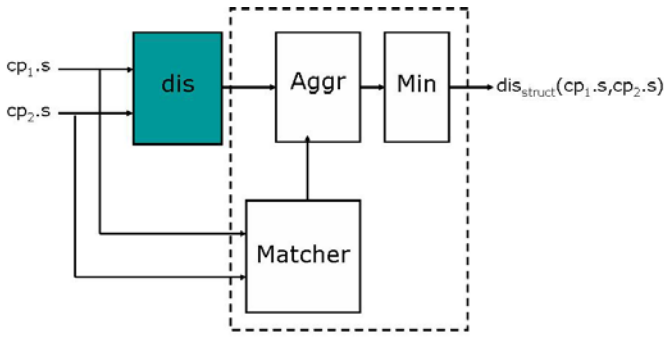
συναθροιστούν έτσι ώστε να προκύψει μία μοναδική, συνολική τιμή ανομοιότητας. Γενικά, μία τέτοια συνάθροιση επιτυγχάνεται μέσω μίας συνάρτησης $Aggr$, η οποία παίρνει ως είσοδο τον πίνακα $D = (dis(p_1^i, p_2^j))$ των ανομοιοτήτων των επιμέρους προτύπων και έναν πίνακα ταιριασμάτων X, δηλαδή,

$$Aggr(D, X)$$

Από όλα τα έγκυρα ταιριάσματα (όπου η εγκυρότητα καθορίζεται από τον τύπο ταιριάσματος), η λογική είναι να επιλεγεί το καλύτερο, δηλαδή, εκείνο που ελαχιστοποιεί την ολική τιμή της ανομοιότητας:

$$dis_{struct}(cp1.s, cp2.s) = \min_X \{Aggr(D, X)\} \tag{3.11}$$

Η διαδικασία που ακολουθείται από το $PANDA$ για να υπολογιστεί η δομική ανομοιότητα μεταξύ δύο σύνθετων προτύπων συνοψίζεται στο Σχήμα 3.6 (αντιστοιχεί στο έγχρωμο κουτί του Σχήματος 3.1).



Σχήμα 3.6: Αποτίμηση της δομικής ανομοιότητας μεταξύ σύνθετων προτύπων

Η ιδέα είναι η ακόλουθη:

- Το τμήμα *Ταίριασμα (Matcher)* παράγει όλα τα δυνατά ταιριάσματα μεταξύ των επιμέρους προτύπων των προς σύγκριση σύνθετων προτύπων. Προφανώς, οι περιορισμοί που επιβάλλονται από τον τύπο ταιριάσματος λαμβάνονται υπόψη.

- Το τμήμα *Συνάθροιση (Aggr)* συναθροίζει τις τιμές των ταιριασμένων επιμέρους προτύπων, υπολογίζοντας μία ολική τιμή για το συγκεκριμένο ταίριασμα.

- Το τμήμα *Ελαχιστοποίηση (Min)* εξετάζει όλα τα ταιριάσματα, και επιλέγει τα καλύτερο, δηλαδή, εκείνο το οποίο παράγει την μικρότερη τιμή ανομοιότητας.

Στην περίπτωση της συνάθροισης πολλαπλών επιπέδων, το τμήμα ανομοιότητας (το χρωματιστό κουτί του Σχήματος 3.6) μπορεί να περικλείει τον αναδρομικό υπολογισμό της ανομοιότητας μεταξύ σύνθετων προτύπων.

Τελικά, η ολική ανομοιότητα μεταξύ των cp_1, cp_2 είναι όπως στην Εξίσωση 3.1, και ακολουθεί την ίδια λογική όπως στην περίπτωση των απλών προτύπων.

Θα πρέπει να σημειωθεί ότι το Σχήμα 3.6 παρουσιάζει μόνο τα βασικά στοιχεία του πλαισίου μας για την ευκολότερη παρουσίαση των επιχειρημάτων μας. Στην πραγματικότητα, δεν απαιτούμε η ανομοιότητα να υπολογίζεται με αυτόν τον τρόπο. Πιο συγκεκριμένα, το τμήμα "Ταίριασμα" δεν χρειάζεται να παράγει όλα τα έγκυρα ταιριάσματα (αυτό θα ήταν μη αποδοτικό στην πράξη). Πράγματι, αποδοτικοί αλγόριθμοι ταιριασμάτων μπορούν να επινοηθούν με σκοπό τον γρήγορο υπολογισμό, δοθέντων των περιορισμών, της λύσης για το πρόβλημα του καλύτερου ταιριάσματος (αυτό αναπαρίσταται από το κουτί με τις διακεκομμένες γραμμές του Σχήματος 3.6). Για παράδειγμα, για τον 1–1 τύπο ταιριάσματος, που συμπίπτει με το γνωστό πρόβλημα της ανάθεσης στην θεωρεία γράφων , η βέλτιστη λύση δίνεται από τον Ουγγρικό αλγόριθμο [41], ο οποίος υλοποιεί και το Aggr και το Min τμήμα του Σχήματος 3.6. Για το ίδιο πρόβλημα, ωστόσο, κάποιος θα μπορούσε να χρησιμοποιήσει κάποιον Άπληστο (Greedy) αλγόριθμο (ο οποίος, σε κάθε βήμα, επιλέγει την πιο επικερδή ανάθεση μεταξύ δύο επιμέρους προτύπων) που υλοποιεί μόνο το Aggr τμήμα του Σχήματος 3.6 και δεν προσφέρει τη βέλτιστη λύση. Ο Ουγγρικός αλγόριθμος με πολυπλοκότητα $O(n^3)$ είναι πιο δαπανηρός από τον Άπληστο αλγόριθμο, ο οποίος έχει πολυπλοκότητα $O(n^2)$. Ο Ουγγρικός αλγόριθμος ωστόσο, παρέχει τη βέλτιστη λύση στο πρόβλημα της ανάθεσης σε αντίθεση με τον Άπληστο αλγόριθμο που μπορεί να κολλήσει σε κάποιο τοπικό βέλτιστο.

Παρακάτω, παρουσιάζουμε ένα παράδειγμα σύγκρισης σύνθετων προτύπων. Πιο συγκεκριμένα, θεωρούμε το πρόβλημα της σύγκρισης δικτυακών τόπων. Όπως έχουμε ήδη αναφέρει, ένας δικτυακός τόπος είναι ένα σύνθετο πρότυπο που αποτελείται από δικτυακές σελίδες, οι οποίες με τη σειρά τους είναι επίσης σύνθετα πρότυπα που αναπαρίστανται ως σύνολα από λέξεις κλειδιά. Οι λέξεις κλειδιά τελικώς, παίζουν το ρόλο των απλών προτύπων. Ξεκινάμε με τη σύγκριση δικτυακών σελίδων και στη συνέχεια προχωράμε στη σύγκριση δικτυακών τόπων. Η σύγκριση μεταξύ λέξεων κλειδιών θα μπορούσε να πραγματοποιηθεί όπως έχουμε ήδη περιγράψει στο Παράδειγμα 10.

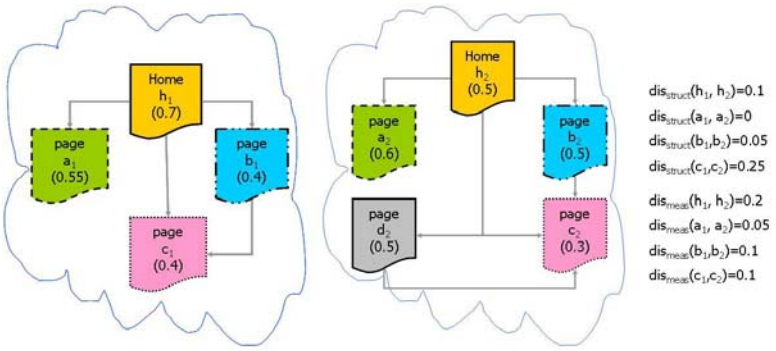
Παράδειγμα 11 (Δικτυακές σελίδες) *Μία δικτυακή σελίδα wp μπορεί να μοντελοποιηθεί ως ένα σύνθετο πρότυπο του οποίου η δομή αποτελείται από ένα σύνολο από λέξεις κλειδιά, $wp.s = \{(t^1, w^1), \ldots, (t^M, w^M)\}$. Το μέτρο της wp μπορεί να είναι το Google PageRank της [13], $wp.m = pr$. Η ανομοιότητα μεταξύ δύο δικτυακών σελίδων, wp_1 και wp_2, θα μπορούσε να λάβει υπόψη μόνο τις λέξεις*

κλειδιά τους, ή θα μπορούσε επίσης να θεωρήσει και τις διαφορές στα PageRank μέτρα τους. Όσο για τη συνάρτηση συνάθροισης, θα μπορούσαμε να χρησιμοποιήσουμε την αλγεβρική συνένωση των τιμών της δομικής και της ποσοτικής ανομοιότητας. Πιο συγκεκριμένα, για τη δομική ανομοιότητα των wp_1, wp_2 κάποιος θα μπορούσε να εφαρμόσει το 1 - 1 ταίριασμα μεταξύ των επιμέρους λέξεων - κλειδιά και τη συνάρτηση μέσου όρου (avg) για τη συνάθροιση των τιμών των ταιριασμένων ζευγών.

□

Παράδειγμα 12 (Δικτυακοί τόποι) *Συνεχίζοντας το Παράδειγμα 11, ένας δικτυακός τόπος ws μπορεί να μοντελοποιηθεί ως ένα σύνθετο πρότυπο του οποίου η δομή αποτελείται από έναν γράφο, δηλαδή, ένα σύνολο από σελίδες $\{wp^1, wp^2, \ldots, wp^N\}$ οι οποίες είναι συνδεδεμένες μέσω συνδέσμων της μορφής $(wp^i \rightarrow wp^j)$. Η ανομοιότητα μεταξύ δύο δικτυακών τόπων, ws_1 και ws_2, μπορεί να υπολογιστεί θεωρώντας το πρόβλημα του ταιριάσματος υπο-γράφων [79], δηλαδή, βρίσκοντας κατά πόσο υπάρχει ισομορφισμός μεταξύ (υπογράφων) των ws_1 και ws_2.*

Ως παράδειγμα, ας θεωρήσουμε τους δικτυακούς τόπους του Σχήματος 3.7, όπου οι σελίδες του ιδίου χρώματος και περιθωρίου αναφέρονται στο ίδιο θέμα.



Σχήμα 3.7: Δύο δικτυακοί τόποι (σελίδες του ιδίου χρώματος και περιγράμματος αναφέρονται στο ίδιο θέμα) και οι τιμές της δομικής και ποσοτικής ανομοιότητας μεταξύ των επιμέρους ταιριασμένων σελίδων τους.

Έστω ότι η ανομοιότητα μεταξύ των δικτυακών τόπων υπολογίζεται όπως στο Παράδειγμα 11 και οι ταιριασμένες σελίδες είναι εκείνες που απεικονίζονται με το ίδιο χρώμα και περίγραμμα στο Σχήμα 3.7. Έστω επίσης ότι οι τιμές της δομικής και της ποσοτικής ανομοιότητας μεταξύ των ταιριασμένων σελίδων φαίνονται στο δεξί τμήμα του σχήματος αυτού. Αν επιπλέον, η συνάρτηση συνάθροισης είναι ο μέσος όρος, τότε η ολική τιμή της ανομοιότητας μεταξύ των δύο δικτυακών τόπων

υπολογίζεται ως ακολούθως:

$$dis(ws_1, ws_2) = dis_{struct}(ws_1.s, ws_2.s) =$$
$$\frac{dis(h_1, h_2) + dis(a_1, a_2) + dis(b_1, b_2) + dis(c_1, c_2) + dis(\perp, d_2)}{5} =$$
$$\frac{(0.1 + 0.2 - 0.1 * 0.2) + (0 + 0.05 - 0 * 0.05) + (0.05 + 0.1 - 0.05 * 0.1)}{5} +$$
$$\frac{(0.25 + 0.1 - 0.25 * 0.1) + (1)}{5} =$$
$$\frac{1.8}{5} = 0.36$$

□

Τέλος, σημειώνουμε πως παρότι η σύγκριση μεταξύ προτύπων επιστρέφει μία απλή τιμή ανομοιότητας, το *PANDA* μπορεί επίσης να παρέχει λεπτομερείς πληροφορίες για το πώς τα επιμέρους πρότυπα ταίριαξαν, όπως και για το πόσο σημαντικά είναι αυτά τα ταιριάσματα στη συνολική τιμή της ανομοιότητας. Με τον τρόπο αυτό, το *PANDA* επιτρέπει στον τελικό χρήστη να αυξήσει την γνώση του για το αποτέλεσμα της σύγκρισης, π.χ., να καταλάβει ποια είναι τα επιμέρους πρότυπα που συνεισφέρουν περισσότερο στην αύξηση της τιμής της ανομοιότητας.

3.4 Θέματα υλοποίησης

Το πλαίσιο *PANDA* έχει υλοποιηθεί σε Java και πολλές ενδεικτικές εφαρμογές για τη σύγκριση διαφορετικών τύπων προτύπων έχουν αναπτυχθεί (Ο κώδικας είναι ελεύθερα διαθέσιμος για μη εμπορική χρήση. Οι ενδιαφερόμενοι χρήστες μπορούν να τον κατεβάσουν από εδώ: http://195.251.230.17/panda/index.html(Ενότητα 3.4.1) και στη συνέχεια περιγράφουμε εν συντομία την εφαρμογή (Ενότητα 3.4.2).

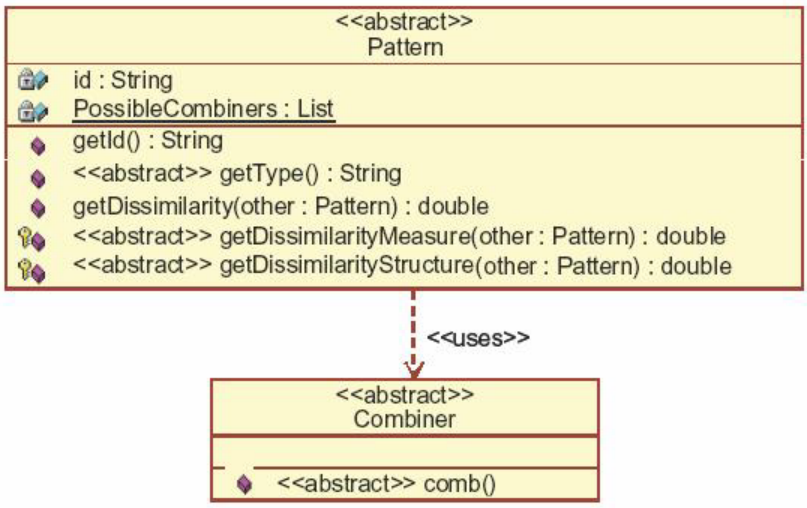
3.4.1 Βασικές κλάσεις του πλαισίου

Οι βασικές κλάσεις είναι η γενική κλάση *Pattern* και οι υποκλάσεις της: *SimplePattern* κλάση και *ComplexPattern* κλάση.

Η κλάση Pattern: Ο πυρήνας της υλοποίησης του πλαισίου είναι η αφηρημένη κλάση *Pattern* (Σχήμα 3.8). Κάθε pattern έχει ένα μοναδικό αναγνωριστή (unique identifier), *id* και ανήκει σε ένα συγκεκριμένο τύπο προτύπου, που μπορεί να επιστραφεί μέσω της μεθόδου *getType()*.

Η μέθοδος *getDissimilarityStructure()* υπολογίζει τη δομική ανομοιότητα μεταξύ του συγκεκριμένου προτύπου και ενός άλλου προτύπου που λαμβάνεται ως είσοδος. Και τα δύο πρότυπα θα πρέπει να ανήκουν στον ίδιο τύπο προτύπων. Αντίστοιχα, υπάρχει η μέθοδος *getDissimilarityMeasure()* για την εκτίμηση της μετρικής ανομοιότητας μεταξύ του συγκεκριμένου προτύπου και ενός δοθέντος ως είσοδο προτύπου. Η μέθοδος *getDissimilarity()* υλοποιεί την συνάρτηση συνδυασμού, *Comb*, δηλαδή, συνδυάζει τις τιμές της δομικής και της μετρικής ανομοιότητας σε μία ολική τιμή ανομοιότητας.

Οι μέθοδοι *getType()*, *getDissimilarityStructure()*, *getDissimilarityMeasure()* και *getDissimilarity()* είναι αφηρημένες και θα πρέπει να υλοποιούνται από κάθε κλάση που επεκτείνει την βασική κλάση *Pattern*. Η μέθοδος *getDissimilarity()* χρησιμοποιεί ένα αντικείμενο της αφηρημένης κλάσης *Combiner* που αντιστοιχεί στις διαφορετικές συναρτήσεις συνδυασμού των δομικών και ποσοτικών ανομοιοτήτων.



Σχήμα 3.8: Η κλάση *Pattern*

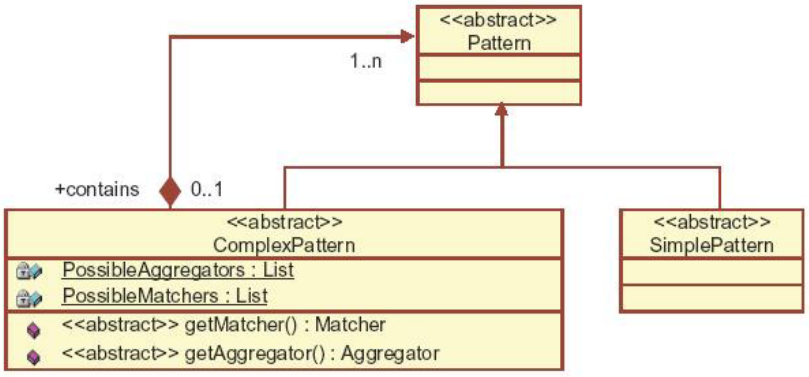
Η κλάση SimplePattern: Η κλάση *SimplePattern* υλοποιεί ένα στιγμιότυπο ενός προτύπου του τύπου των απλών προτύπων και επεκτείνει την βασική κλάση *Pattern* (Σχήμα 3.9).

Οι διάφορες συναρτήσεις δομικής ανομοιότητας που είναι διαθέσιμες για κάθε τύπο προτύπου αποθηκεύονται στο στατικό μέλος *PossibleStructureDissimilarity*. Διάφορες συναρτήσεις δομικής ανομοιότητας έχουν υλοποιηθεί μέσα στο package *dissimilarity.structure*, όπως η *EuclideanDistance* και η *JaccardDistance*.

Οι διάφορες συναρτήσεις ποσοτικής ανομοιότητας που είναι διαθέσιμες για κάθε τύπο προτύπου αποθηκεύονται στο στατικό μέλος (static member) *PossibleMeasureDissimilarity*. Διάφορες συναρτήσεις μετρικής ανομοιότητας έχουν υλοποιηθεί μέσα στο package *dissimilarity.measure* όπως η *AbsoluteDistance* και η *RelativeDistance*.

Οι διάφορες συναρτήσεις συνδυασμού που είναι διαθέσιμες για κάθε τύπο προτύπου αποθηκεύονται στο στατικό μέλος*PossibleCombiners*. Διάφορες συναρτήσεις συνδυασμού έχουν υλοποιηθεί μέσα στο package *dissimilarity.combiner* όπως η *CombinerStructure* η οποία χρησιμοποιεί μόνο την τιμή της δομικής ανομοιότητας και η *CombinerWeighted* που λαμβάνει υπόψη τις τιμές της δομικ•ση και της μετρικής ανομοιότητας σταθμισμένες ανάλογα με τις επιλογές του χρήστη.

Η κλάση ComplexPattern: Η κλάση *ComplexPattern* υλοποιεί ένα στιγμιότυπο προτύπου με τύπο σύνθετου προτύπου και επεκτείνει την βασική κλάση



Σχήμα 3.9: Η ιεραρχία των κλάσεων *Pattern*

Pattern (Σχήμα 3.9). Ένα σύνθετο πρότυπο αποτελείται από μία λίστα από απλά πρότυπα, αποθηκευμένα στο μέλος *PatternList*.

Οι διαφορετικοί τύποι ταιριάσματος που είναι διαθέσιμοι για κάθε τύπο προτύπου αποθηκεύονται στο στατικό μέλος *PossibleMatchers*, το οποίο περιέχει αναφορές στα αντικείμενα *Matcher*. Διάφοροι ταιριαστές (matchers) έχουν υλοποιηθεί μέσα στο package *dissimilarity.matcher* όπως ο *MatcherHungarian*, ο *MatcherGreedy* και ο *MatcherMN*.

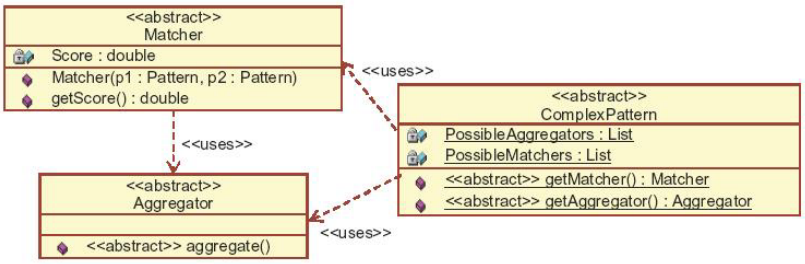
Οι διάφορες συναρτήσεις συνάθροισης που είναι διαθέσιμες για κάθε τύπο προτύπων αποθηκεύονται στο στατικό μέλος *PossibleAggregators*, το οποίο περιέχει αναφορές σε αντικείμενα *Aggregator*. Διάφορες συναρτήσεις συνάθροισης έχουν υλοποιηθεί μέσα στο package *dissimilarity.aggregator*, όπως η *AggregatorSimpleAvg* η οποία υπολογίζει τη μέση τιμή των ανομοιοτήτων των ταιριασμένων επιμέρους προτύπων, η *AggregatorMin* η οποία υπολογίζει την ελάχιστη τιμή των ανομοιοτήτων των ταιριασμένων επιμέρους προτύπων και η *AggregatorMax* η οποία υπολογίζει τη μέγιστη από τις ανομοιότητες των ταιριασμένων επιμέρους προτύπων.

Οι συσχετίσεις μεταξύ σύνθετων προτύπων, τύπος ταιριασμάτων και συναρτήσεων συνάθροισης παρουσιάζονται στο Σχήμα 3.10, όπου διευκρινίζεται ότι κάθε τύπος προτύπων χρησιμοποιεί τα δικά του *Matcher* και *Aggregator* αντικείμενα, που έχουν επιλεγεί από τις λίστες *PossibleMatchers* και *PossibleAggregators*, αντίστοιχα.

3.4.2 Η εφαρμογή

Η εφαρμογή μας επιτρέπει τη σύγκριση διαφορετικών τύπων προτύπων, και συγκεκριμένα συχνών στοιχειοσυνόλων, συστάδων, χρονοσειρών, συνόλων από κείμενα κλπ. Ακολούθως, καταγράφουμε λεπτομερώς την υλοποίηση του *PANDA* για τη σύγκριση μηνιαίων χρονοσειρών από δεδομένα που λήφθησαν από το Ιταλικό MIB [83]. Επιλέξαμε χρονοσειρές ως παράδειγμα καθώς η οπτικοποίηση των αποτελεσμάτων είναι ευκολότερη σε σχέση με άλλους τύπους, όπως για παράδειγμα τα συχνά στοιχειοσύνολα.

Για να μπορεί το πλαίσιο *PANDA* να εφαρμοσθεί σε κάποιο συγκεκριμένο πρόβλημα σύγκρισης προτύπων, θα πρέπει τα πρότυπα που θα συγκριθούν να ανα-



Σχήμα 3.10: Σύνθετα πρότυπα, τύποι ταιριάσματος και συναρτήσεις συνάθροισης

παρασταθούν σύμφωνα με τη λογική των απλών-σύνθετων προτύπων. Στη συνέχεια, για την περίπτωση των απλών προτύπων, θα πρέπει να ορισθούν οι *getDissimilarityStructure()*, *getDissimilarityMeasure()* και *getDissimilarity()* μέθοδοι, ενώ για την περίπτωση των σύνθετων προτύπων, θα πρέπει να ορισθούν οι συναρτήσεις *Matcher* και *Aggregator*. Ακολούθως, περιγράφουμε τα βήματα αυτά για την περίπτωση της εφαρμογή μας.

3.4.2.1 Ορισμός των τύπων προτύπων

Ο τύπος προτύπων $StockValue$ είναι ένας *απλός* τύπος προτύπων που ορίζεται ως εξής:

$$\begin{aligned}&\texttt{StockValue} = \\ &\qquad (SS : \texttt{(month:Integer, year:Integer)}, \\ &\qquad MS : (\texttt{[Real]}_1^N))\end{aligned}$$

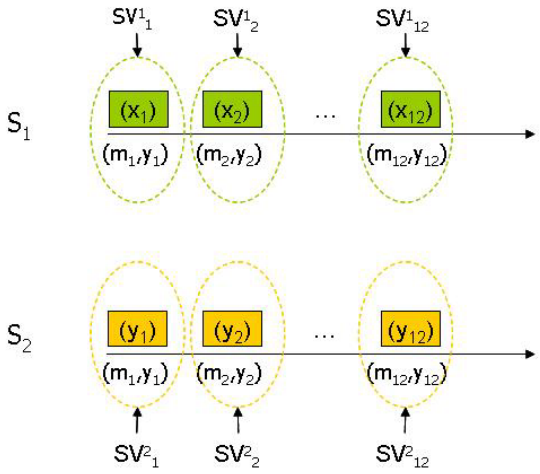
Ο τύπος προτύπων *Stock* είναι ένα *σύνθετο* τύπος προτύπων που αποτελείται από $StockValue$ αντικείμενα:

$$\begin{aligned}&\texttt{Stock} = \\ &\quad (SS : \{\texttt{TimeSeries}\}, \\ &\quad MS : \bot)\end{aligned}$$

Ένα παράδειγμα δύο *Stock* προτύπων και τα επιμέρους τους $StockValue$ πρότυπα παρουσιάζεται στο Σχήμα 3.11):

Ένας $SetOfStocks$ τύπος προτύπων είναι ένας *σύνθετος* τύπος προτύπων που αποτελείται από *Stock* πρότυπα:

$$\begin{aligned}&\texttt{SetOfStocks} = \\ &\qquad (SS : \{\texttt{Stock}\}, \\ &\qquad MS : \bot)\end{aligned}$$



Σχήμα 3.11: Δύο *Stock* πρότυπα (συμβολίζονται με S) και τα επιμέρους $StockValue$ πρότυπα (συμβολίζονται με SV)

Σύγκριση μεταξύ (απλών) $StockValue$ **προτύπων:** Η δομική ανομοιότητα, dis_{struct}, ισούται με 1, αν η δομές είναι διαφορετικές, αλλιώς ισούται με 0. Η ποσοτική ανομοιότητα, dis_{meas}, ισούται με την απόλυτη διαφορά των (κανονικοποιημένων) μέτρων τους. Οι τιμές της δομικής και της ποσοτικής ανομοιότητας συνδυάζονται (συνάρτηση $Comb$) έτσι ώστε αν $dis_{struct} = 1$, η ολική απόσταση dis να είναι επίσης 1, αλλιώς $dis = dis_{meas}$.

Σύγκριση μεταξύ (σύνθετων) *Stock* **προτύπων:** Η ολική ανομοιότητα μεταξύ δύο *Stock* προτύπων λαμβάνεται από το μέσο όρο των ανομοιοτήτων των επιμέρους $StockValue$ προτύπων του ίδιου μήνα. Συνεπώς, το τμήμα *Ταιριαστής* αρχικοποιείται με τύπο ταιριάσματος 1–1, ενώ το τμήμα *Συναθροιστής* αρχικοποιείται με τη συνάρτηση μέσου όρου.

Σύγκριση μεταξύ (σύνθετων) $SetOfStocks$ **προτύπων:** Η ολική ανομοιότητα μεταξύ δύο $SetOfStocks$ λαμβάνεται από το μέσο όρο των ανομοιοτήτων μεταξύ των επιμέρους *Stock* προτύπων. Ξανά, το 1–1 ταίριασμα επιλέγεται ως *Matcher* μεταξύ των επιμέρους *Stock* προτύπων και η συνάρτηση μέσης τιμής επιλέγεται να είναι ο *Aggregator*.

Ορισμός του τύπου ταιριάσματος: Όπως έχουμε ήδη δηλώσει, σε αυτή την ενδεικτική εφαρμογή εφαρμόζουμε ένα 1–1 ταίριασμα μεταξύ των επιμέρους προτύπων. Πειραματισθήκαμε με δύο διαφορετικούς αλγορίθμους 1–1 ταιριάσματος: το βέλτιστο *HungarianMatcher* αλγόριθμο και τον *GreedyMatcher* αλγόριθμο. Ο πρώτος βρίσκει το βέλτιστο 1–1 ταίριασμα μεταξύ των επιμέρους προτύπων βασισμένος στον Ουγγρικό αλγόριθμο [41], ενώ ο δεύτερος, ακολουθώντας την προσέγγιση των "άπληστων" αλγορίθμων, ταιριάζει σε κάθε βήμα τα λιγότερο ανόμοια μη ταιριασμένα επιμέρους πρότυπα. Ο άπληστος αλγόριθμος είναι γρηγορότερος από τον Ουγγρικό, ωστόσο μπορεί να κολλήσει σε κάποιο τοπικό βέλτιστο, δηλαδή, δεν εξασφαλίζει ότι θα βρει την βέλτιστη λύση.

Ορισμός της λογικής συνάθροισης: Σε αυτή την εφαρμογή, χρησιμοποιούμε την μέση τιμή ως συνάρτηση συνάθροισης.

Το περιβάλλον διεπαφής της εφαρμογής: Η βασική φόρμα της εφαρμογής φαίνεται στο Σχήμα 3.12. Στην αριστερά πλευρά του παραθύρου, ο χρήστης μπορεί να επιλέξει, μεταξύ των διαθέσιμων τύπων προτύπων, τον (σύνθετο) τύπο προτύπων που επιθυμεί να χρησιμοποιήσει. Μετά την επιλογή του τύπου προτύπων, εμφανίζονται η λίστα των διαθέσιμων τύπων ταιριάσματος και των συναρτήσεων συνάθροισης για τον τύπο αυτό και ο χρήστης μπορεί να επιλέξει τον τύπο ταιριάσματος και τον συναθροιστή για την εφαρμογή του.

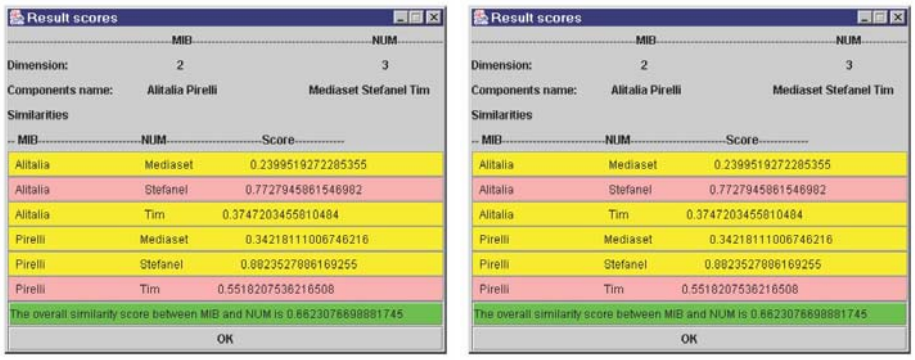
Σχήμα 3.12: Βασική φόρμα της εφαρμογής

Στο Σχήμα 3.14 παρουσιάζουμε τα αποτελέσματα της σύγκρισης δύο ενδεικτικών $SetOfStocks$ προτύπων, δηλαδή τα πρότυπα "MIB" και "NUM". Το πρώτο σύνολο αποτελείται από 2 *stocks* ("Alitalia", "Pirelli"), ενώ το δεύτερο αποτελείται από 3 *stocks* ("Mediaset", "Stefanel", "TIM"). Το παράθυρο εμφανίζει τις τιμές (εδώ παριστάνει τιμές ομοιότητας υπολογισμένες ως $1-dis$) για κάθε ζεύγος μετοχών· τα καλύτερα ταιριάσματα τονίζονται με ροζ χρώμα. Τέλος, η ολική τιμή φαίνεται με πράσινο χρώμα στο κάτω μέρος του παραθύρου. Τα αποτελέσματα που λάβαμε μέσω του *HungarianMatcher* παρουσιάζονται στο αριστερό μέρος του σχήματος, ενώ εκείνα που λάβαμε με χρήση του *GreedyMatcher* φαίνονται στο δεξιό τμήμα του σχήματος. Όπως γίνεται προφανές από αυτό το σχήμα, η τιμή της ανομοιότητας που δίνει ο Ουγγρικός αλγόριθμος είναι χαμηλότερη από την τιμή που δίνει ο Άπληστος αλγόριθμος, μία συνέπεια του γεγονότος ότι ο Άπληστος αλγόριθμος μπορεί να κολήσει σε κάποιο τοπικό ακρότατο, ενώ ο Ουγγρικός εγγυάται την βέλτιστη λύση.

Οι ταιριασμένες χρονοσειρές μπορούν επίσης να παρουσιαστούν γραφικά στην (3.15), όπου οι ταιριασμένες μετοχές φαίνονται στην ίδια γραμμή μαζί με την τιμή της ομοιότητας, ενώ οι μη ταιριασμένες μετοχές παρουσιάζονται μόνες.

3.5 Εφαρμογές του *PANDA* σε διάφορους τύπους προτύπων

Εφαρμόσαμε το *PANDA* σε διάφορες εφαρμογές σύγκρισης προτύπων προκειμένου να δείξουμε την χρησιμότητά του σε ένα ευρύ φάσμα εφαρμογών.



(α) τον Ουγγρικό αλγόριθμο (β) τον Άπληστο αλγόριθμο

Σχήμα 3.13: Σύγκριση δύο $SetOfStocks$ προτύπων μέσω 1–1 ταιριάσματος χρησιμοποιώντας

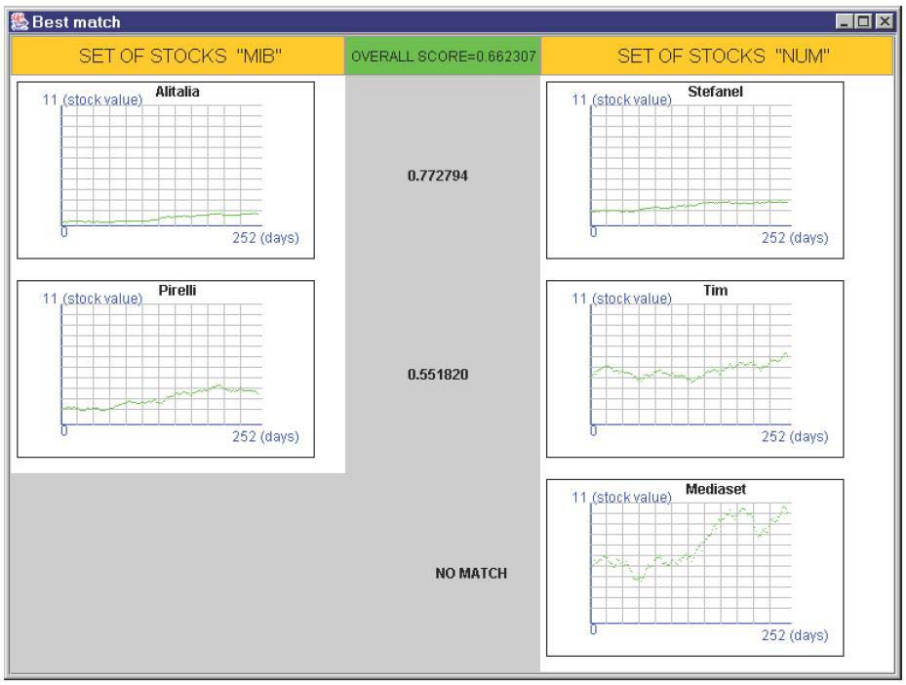
Σχήμα 3.14: Σύγκριση δύο $SetOfStocks$ προτύπων χρησιμοποιώντας τον Ουγγρικό αλγόριθμο (αριστερά) και τον Άπληστο αλγόριθμο (δεξιά)

3.5.1 Εφαρμογή σε σύνολα από συχνά στοιχειοσύνολα

Στο πείραμα αυτό χρησιμοποιήσαμε τον γεννήτορα [5], ο οποίος γεννά δεδομένα που μιμούνται τις συναλλαγές των πελατών σε ένα κατάστημα λιανικής, για να δημιουργήσουμε ένα σύνολο δεδομένων D που αποτελείται από 1000 συναλλαγές με μέσο μήκος συναλλαγής 10. Για την εξαγωγή των συχνών στοιχειοσυνόλων χρησιμοποιήσαμε το πρόγραμμα $MAFIA$ [14].

Με βάση το μοντέλο αναπαράστασης, ένα στοιχειοσύνολο ($Itemset$) μπορεί να μοντελοποιηθεί ως ένα απλό πρότυπο $(s = S, m = supp)$ όπου S είναι το σύνολο των στοιχείων που το αποτελούν (δομική συνιστώσα) και $supp \in [0, 1]$ είναι η υποστήριξη του στοιχειοσυνόλου (ποσοτική συνιστώσα), βλέπε επίσης Παράδειγμα 9. Ένα σύνολο από στοιχειοσύνολα ($SetOfItemsets$) μπορεί να μοντελοποιηθεί ως ένα σύνθετο πρότυπο, που αποτελείται από (απλά) πρότυπα τύπου $Itemset$.

Προκειμένου να συγκρίνουμε τα πρότυπα τύπου $Itemset$ (απλά πρότυπα) θα πρέπει να ορίσουμε πως συγκρίνονται οι δομικές και οι ποσοτικές τους συνιστώσες και πως τα επιμέρους σκορ τους συνδυάζονται. Για τη συνάρτηση της ποσοτικής



Σχήμα 3.15: Οπτικοποίηση των ταιριασμάτων μεταξύ των δύο $\Sigma\epsilon\tau O\varphi\Sigma\tau o\varsigma\kappa\varsigma$ προτύπων του Σχήματος 3.14

ανομοιότητας, χρησιμοποιήσαμε την απόλυτη διαφορά των μέτρων τους, δηλαδή:

$$dis_{meas}(p^i, p^j) = |p^i.m - p^j.m|$$

Για τη συνάρτηση της δομικής ανομοιότητας, χρησιμοποιήσαμε την επικάλυψη των στοιχείων που συνθέτουν κάθε στοιχειοσύνολο:

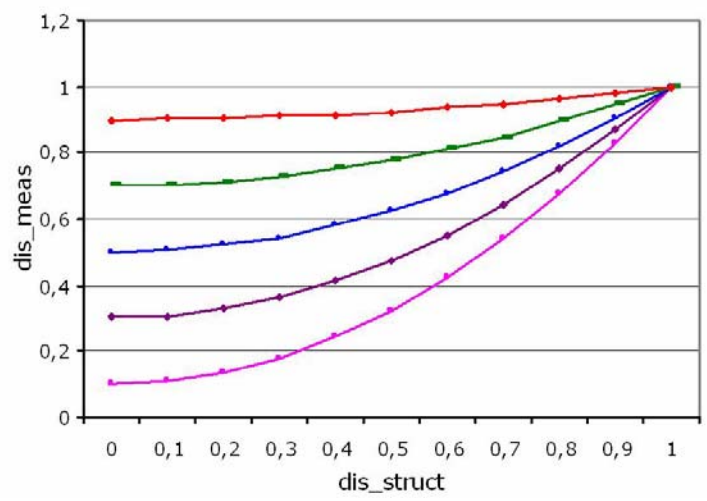
$$dis_{struct}(p^i, p^j) = \frac{|p^i.s \cap p^j.s|}{|p^i.s \cup p^j.s|}$$

Για τη συνάρτηση συνδυασμού, χρησιμοποιήσαμε την ακόλουθη συνάρτηση:

$$Comb(p^i, p^j) = dis_{struct}(p^i, p^j) + \left(1 - dis_{struct}(p^i, p^j)\right) * dis_{meas}(p^i, p^j)^2$$

Η λογική πίσω από αυτή τη συνάρτηση είναι ότι όσο πιο όμοια είναι τα δύο στοιχειοσύνολα όσον αφορά στη δομή τους, δηλαδή, όσο πιο χαμηλό το dis_{struct} σκορ, τόσο περισσότερο θα πρέπει να ληφθούν υπόψη στον υπολογισμό του τελικού σκορ οι ποσοτικές συνιστώσες τους, μέσω του σκορ dis_{meas}. Αυτή η συμπεριφορά απεικονίζεται στην Εικόνα 3.16.

Προκειμένου να συγκρίνουμε πρότυπα τύπου $SetOfitemset$ θα πρέπει να ορίσουμε πως τα επιμέρους $Itemset$ πρότυπα που τα αποτελούν θα ταιριάξουν και πως τα σκορ των ταιριασμάτων θα συναθροιστούν. Για το ταίριασμα, χρησιμοποιούμε τον 1-1 τύπο ταιριάσματος, ενώ για τη λογική συνάθροισης, χρησιμοποιούμε τον μέσο όρο. Όσον αφορά στο 1–1 ταίριασμα πειραματιστήκαμε τόσο με τον Ουγγρικό αλγόριθμο [41] όσο και με τον Άπληστο αλγόριθμο (βλέπε Ενότητα 3.3.2)



Σχήμα 3.16: Η συμπεριφορά του συνδυαστή ως προς τα dis_{struct} και dis_{meas}

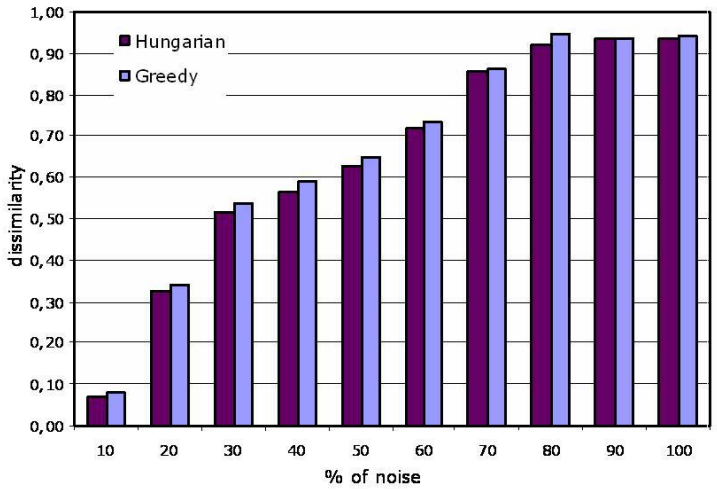
Στη συνέχεια τρέξαμε ένα πείραμα όπου προσθέταμε διαφορετικά επίπεδα θορύβου στο αρχικό σύνολο δεδομένων $D_0 \equiv D$: Συγκεκριμένα, σε κάθε βήμα εισάγαμε διαφορετικά επίπεδα θορύβου στο D_0, δηλαδή, 10%, 20%, ..., 100%. Σε κάθε συναλλαγή που "πειράξαμε", αλλάξαμε ένα συγκεκριμένο ποσοστό (50%)των στοιχείων της. Από τα σύνολα δεδομένων που προέκυψαν (D_0, D_1, ..., D_{10}), εξάγαμε τα αντίστοιχα σύνολα από στοιχειοσύνολα $(P_0, P_1, \ldots, P_{10})$, κάτω από $minSupport$=20%. Στη συνέχεια, συγκρίναμε τα "πειραγμένα" σύνολα προτύπων P_i $(i = 1, 2, \ldots, 10)$, με το αρχικό "καθαρό" σύνολο προτύπων P_0. Ο στόχος ήταν να δούμε πως ο θόρυβος που προσθέταμε στα δεδομένα επηρεάζει τά εξαγόμενα πρότυπα.

Όπως φαίνεται στην Εικόνα 3.17, η ανομοιότητα στο χώρο των προτύπων αντανακλά την αύξηση της ανομοιότητας στο χώρο των πρωτογενών δεδομένων λόγω της προσθήκης θορύβου. Η ανομοιότητα αυξάνει σταδιακά καθώς το ποσοστό του θορύβου αυξάνεται, και γίνεται σχεδόν ένα όταν το αρχικό σύνολο δεδομένων γεμίζει με θόρυβο.

Συγκρίνοντας τον Ουγγρικό και τον Άπληστο αλγόριθμο, ο Ουγγρικός αλγόριθμος υπολογίζει ελαφρώς μικρότερη ανομοιότητα σε σχέση με τον Άπληστο αλγόριθμο. Αυτό οφείλεται στο γεγονός ότι ο Ουγγρικός αλγόριθμος παρέχει τη βέλτιστη λύση στο πρόβλημα του 1 - 1 ταιριάσματος, ενώ ο Άπληστος παρέχει μία προσεγγιστική λύση. Ωστόσο, λαμβάνοντας υπόψη ότι ο Άπληστος αλγόριθμος είναι πολύ πιο γρήγορος από τον Ουγγρικό (25 φορές, στα πειράματά μας) ένα πρακτικό συμπέρασμα από αυτή την εφαρμογή είναι ότι ο Άπληστος αλγόριθμος θα μπορούσε να χρησιμοποιηθεί με ασφάλεια για τη σύγκριση συνόλων από στοιχειοσύνολα.

3.5.2 Εφαρμογή σε δέντρα απόφασης

Όπως έχουμε ήδη αναφέρει ένα δέντρο απόφασης τμηματοποιεί το χώρο των γνωρισμάτων σε ένα σύνολο μη-επικαλυπτόμενων περιοχών, μέσω των κόμβων φύλλ-



Σχήμα 3.17: Επίδραση του θορύβου των δεδομένων στην ανομοιότητα των αντίστοιχων συνόλων από στοιχειοσύνολα, στην περίπτωση του Ουγγρικού και του Άπληστου αλγορίθμου

ων του. Μία *περιοχή* (region) αποτελεί ένα απλό πρότυπο. Η δομική συνιστώσα μιας περιοχής, αποτελείται από τις συνθήκες ελέγχου στο αντίστοιχο μονοπάτι του δέντρου. Οι συνθήκες αυτές είναι συνήθως αριθμητικές, συνεπώς μπορούν να περιγραφούν ως εξής: $(ValueFrom \leq attribute \leq ValueTo)$, όπου $attribute$ είναι κάποιο γνώρισμα πρόβλεψης του προβλήματος κατηγοριοποίησης και $ValueFrom, ValueTo$ είναι το εύρος τιμών του γνωρίσματος στη συγκεκριμένη περιοχή. Συνεπώς, η *δομική συνιστώσα μιας περιοχής* μπορεί να περιγραφεί ως η συνένωση αυτών των συνθηκών. Η *ποσοτική συνιστώσα μιας περιοχής* αποτελείται από το ποσοστό των στιγμιότυπων του προβλήματος που καταλήγουν στην περιοχή για κάθε κλάση του προβλήματος, $\{(c, n_c)\}$, όπου $c \in C$ είναι η κλάση του προβλήματος και n_c είναι το ποσοστό των στιγμιότυπων που καταλήγουν στη συγκεκριμένη περιοχή και ανήκουν στην κλάση c.

Ένα *δέντρο απόφασης* μπορεί να μοντελοποιηθεί ως σύνθετο πρότυπο όπως παρακάτω:

$$DT = (SS : \{Region\}, \perp)$$

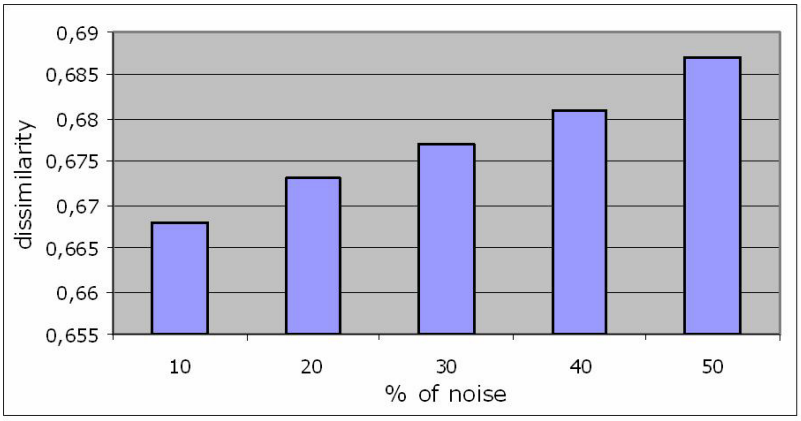
Η ανομοιότητα μεταξύ δύο *περιοχών* (απλά πρότυπα) ορίζεται ως η τομή των αντίστοιχων υπερ-ορθογωνίων τους. Αυτή είναι η δομική ανομοιότητα μεταξύ των περιοχών. Στο πείραμα αυτό, δε λαμβάνουμε υπόψη ξεχωριστά την ποσοτική ανομοιότητα μεταξύ των περιοχών, αλλά θεωρούμε ότι τα στιγμιότυπα του προβλήματος είναι ομοιόμορφα κατανεμημένα στο χώρο γνωρισμάτων και συνεπώς ο όγκος μιας περιοχής αναπαριστά τη σημαντικότητα της περιοχής αυτής σε ολόκληρο το χώρο γνωρισμάτων.

Για να συγκρίνουμε δύο *δέντρα απόφασης* (σύνθετα πρότυπα), χρησιμοποιούμε ένα ταίριασμα τύπου N–M μεταξύ των περιοχών των δέντρων και στη συνέχεια συναθροίζουμε τα σκορ των ταιριασμένων περιοχών χρησιμοποιώντας τη συνάρτηση μέσου όρου (avg).

Για το πείραμά μας, επιλέξαμε το σύνολο δεδομένων $wine$ [11], το οποίο

αποτελείται από τα αποτελέσματα χημικής ανάλυσης 178 κρασιών μιας περιοχής της Ιταλίας.

Εκτελέσαμε το ακόλουθο πείραμα: Πρώτα, κατασκευάσαμε το δέντρο απόφασης DT_0 από το αρχικό σύνολο δεδομένων. Στη συνέχεια προσθέσαμε θόρυβο στα μονοπάτια/ κανόνες που αποτελούν το DT_0 μειώνοντας (αυξάνοντας) κατά $10\%, 20\%, \ldots, 50\%$ την τιμή $ValueFrom$ (αντίστοιχα, $ValueTo$) κάθε γνωρίσματος του κανόνα. Ως αποτέλεσμα, κατασκευάστηκαν τα αντίστοιχα (με θόρυβο) δέντρα απόφασης $(DT_{10}, DT_{20}, \ldots, DT_{50})$. Επίσης, μειώσαμε ή αυξήσαμε τυχαία την υποστήριξη κάθε κανόνα με το αντίστοιχο ποσοστό θορύβου. Μελετήσαμε την επίδραση του θορύβου συγκρίνοντας τα δέντρα με θόρυβο $DT_i, i = \{10, 20, \ldots, 50\}$ με το αρχικό "καθαρό" δέντρο DT_0. Διαισθητικά, όσο πιο πολύ θόρυβο προσθέτουμε, τόσο πιο ανόμοια θα πρέπει να είναι τα δέντρα που προκύπτουν σε σχέση με το αρχικό δέντρο απόφασης. Αυτό επιβεβαιώνεται στην Εικόνα 3.18.



Σχήμα 3.18: Επίδραση του θορύβου (στο χώρο των προτύπων) στην ανομοιότητα μεταξύ δέντρων απόφασης

3.5.3 Εφαρμογή σε συλλογές από κείμενα

Το παρακάτω πείραμα αναφέρεται στη σύγκριση συλλογών από κείμενα. Συγκεκριμένα, χρησιμοποιήσαμε τη βάση δεδομένων $DBLP$ [43], η οποία περιλαμβάνει ένα σύνολο από επιστημονικά περιοδικά. Κάθε περιοδικό περιέχει ένα σύνολο από άρθρα που αναφέρονται στο ίδιο θέμα (το θέμα, δηλαδή που καλύπτεται από το εν λόγω περιοδικό) και κάθε άρθρο περιγράφεται μέσω ενός συνόλου από λέξεις - κλειδιά. Στην περίπτωση αυτή, οι λέξεις - κλειδιά αποτελούν απλά πρότυπα, ενώ τα άρθρα και τα περιοδικά αποτελούν σύνθετα πρότυπα.

Οι λέξεις - κλειδιά (απλά πρότυπα) συγκρίνονται όπως στο Παράδειγμα 10. Για τη σύγκριση των σύνθετων προτύπων (δηλαδή, των άρθρων και των περιοδικών) χρησιμοποιούμε για τον τύπο ταιριάσματος το 1 - 1 ταίριασμα και για τη συνάρτηση συνάθροισης το μέσο όρο.

Τα περιοδικά του $DBLP$ που χρησιμοποιήθηκαν φαίνονται στον Πίνακα 3.2, ενώ τα αποτελέσματα της σύγκρισης φαίνονται στην Εικόνα 3.19.

DBLP Journal	Abbreviation
Computer Journal	Comp J
Artificial Intelligence	AI
Computer Networks	Comp N
Computers & Graphics	Comp G
Information Systems	Info Sys
Computer Networks and ISDN Systems	Comp Net & ISDN Sys
Computational Intelligence	Comp Intell
Computer Languages	Comp Lang
Distributed Computing	Dist Comp
Advances in Computers	Adv in Comp
Evolutionary Computation	Evol Comp
Computational Complexity	Comp Compl
Information Retrieval	IR

Πίνακας 3.2: Τα περιοδικά του $DBLP$

Διάφορα ενδιαφέρονται συμπεράσματα μπορούν να εξαχθούν από το συγκεκριμένο πείραμα:

- Το *Computer Journal* παρουσιάζει τη μεγαλύτερη ομοιότητα (ή μικρότερη ανομοιότητα) με τα υπόλοιπα περιοδικά. Αυτό δικαιολογείται από το γεγονός ότι το εν λόγω περιοδικό παρέχει μία πλήρη επισκόπηση του χώρου της Πληροφορικής.

- Από την άλλη, περιοδικά όπως το *Evolutionary Computation*, *Computational Intelligence*, *Computer Networks*, *ISDN Systems* και *Distributed Computing* παρουσιάζουν μεγάλες ανομοιότητες μεταξύ τους, πιθανόν λόγω του ότι καλύπτουν εξειδικευμένες και διακριτές περιοχές του χώρου της Πληροφορικής.

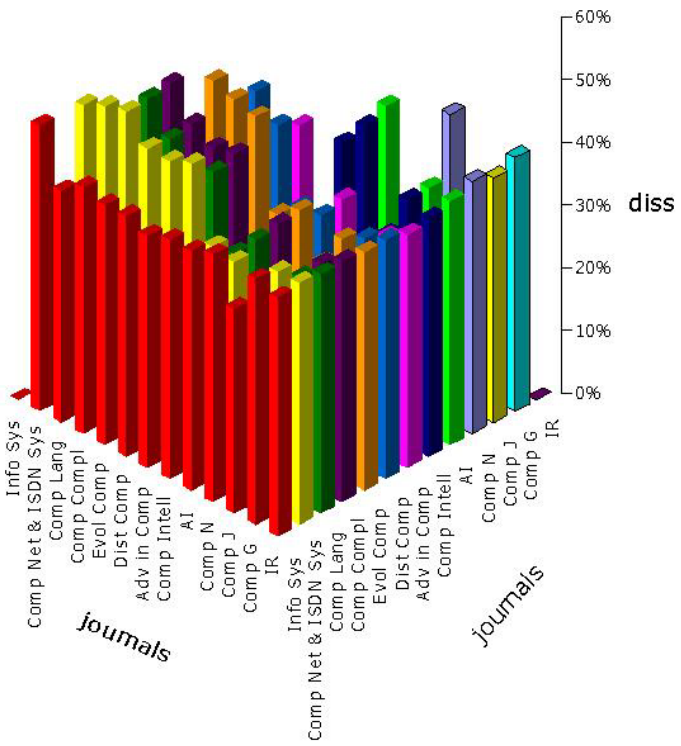
- Η μεγαλύτερη ανομοιότητα βρέθηκε μεταξύ των περιοδικών *Evolutionary Computation* και *Distributed Computing*, τα οποία καλύπτουν εντελώς διακριτές περιοχές.

3.6 Σχετικές εργασίες

Σε αυτή την ενότητα, παρουσιάζουμε περιληπτικά διάφορες προσεγγίσεις που έχουν προταθεί στη βιβλιογραφία και σχετίζονται με το πλαίσιο $PANDA$.

Το πλαίσιο $FOCUS$ Απ΄ όσο γνωρίζουμε, το μόνο γενικό πλαίσιο για τη σύγκριση προτύπων είναι το FOCUS [25], το οποίο αφορά στη σύγκριση συνόλων δεδομένων. Το $FOCUS$ συγκρίνει δύο σύνολα δεδομένων με βάση τα αντίστοιχα σύνολα προτύπων που εξάγονται από τα δεδομένα αυτά. Η λογική πίσω από τη χρήση της σύγκρισης προτύπων για τη σύγκριση συνόλων δεδομένων είναι ότι τα πρότυπα "συλλαμβάνουν" ενδιαφέροντα χαρακτηριστικά των δεδομένων και συνεπώς, μπορούν να χρησιμοποιηθούν για τη σύγκριση των δεδομένων. Τρεις δημοφιλείς τύποι προτύπων, τα συχνά στοιχειοσύνολα, τα δέντρα απόφασης και οι συσταδοποιήσεις, υποστηρίζονται από το $FOCUS$.

Η κεντρική ιδέα του FOCUS, η οποία χρησιμοποιείται επίσης και στο $PANDA$ είναι η μοντελοποίηση των προτύπων μέσω μίας δομικής και μίας ποσοτικής συνιστώσας.



Σχήμα 3.19: Σύγκριση περιοδικών του $DBLP$

Σύμφωνα με το $FOCUS$, ένα πρότυπο p αποτελείται από ένα σύνολο από "περιοχές regions" που ορίζονται στο χώρο των προτύπων (δομική συνιστώσα του p) και κάθε περιοχή συσχετίζεται με ένα σύνολο από μέτρα (ποσοτική συνιστώσα του p).

Τόσο οι δομικές όσο και οι μετρικές συνιστώσες λαμβάνονται υπόψη κατά τη διάρκεια της εκτίμησης της ανομοιότητας. Αν δύο πρότυπα μοιράζονται την ίδια δομή, τότε η ανομοιότητά τους υπολογίζεται συναθροίζοντας (μέσω μίας συνάρτησης g) τις διαφορές στις ποσοτικές συνιστώσες όλων των περιοχών (η διαφορά στις ποσοτικές συνιστώσες μίας περιοχής υπολογίζεται μέσω μία συνάρτησης f). Η συνάρτηση διαφορών f θα μπορούσε να είναι η απόλυτη ή η σχετική διαφορά, ενώ η συνάρτηση συνάθροισης g θα μπορούσε να είναι κάποια από τις sum, max, min κτλπ.. Αν οι δομικές συνιστώσες δεν είναι όμοιες, τότε ένα προκαταρκτικό βήμα χρειάζεται προκειμένου να αναθεωρηθούν. Αυτό περιλαμβάνει τη διάσπαση των περιοχών μέχρι οι δύο δομές να γίνουν όμοιες. Το αποτέλεσμα ονομάζεται Greatest Common Refinement (GCR). Αυτό απαιτεί τον υπολογισμό των μέτρων των (νέων) περιοχών του GCR από τα αρχικά σύνολα δεδομένων. Μετά από αυτό, μπορεί να εφαρμοσθεί ότι έχει περιγραφεί για την βασική περίπτωση των όμοιων δομών.

Συγκρίνοντας το FOCUS με το $PANDA$ προκύπτει ένας αριθμός από περιορισμούς που εμποδίζουν τη χρήση του πρώτου ως ένα γενικό και ευέλικτο πλαίσιο για σύγκριση προτύπων, από την άποψη των απαιτήσεων που παρουσιάσθηκαν στην Ενότητα 3.1:

- Το FOCUS θεωρεί μόνο πρότυπα των οποίων η δομή αποτελείται από μία ιεραρχία δύο επιπέδων (κάθε πρότυπο αποτελείται από περιοχές), ενώ το $PANDA$ υποστηρίζει τον αναδρομικό ορισμό αυθαίρετα σύνθετων προτύπων. Συνεπώς, το $FOCUS$ δεν χειρίζεται ένα πλήθος από ενδιαφέροντες τύπους προτύπων, όπως οι δικτυακοί τόποι του Παραδείγματος 5. Από την άλλη, όλοι οι τύποι προτύπων που υποστηρίζονται από το FOCUS μπορούν να υποστηριχθούν και από το $PANDA$.

- Λόγω της χρήσης του GCR, το FOCUS δεν είναι ευέλικτο ως προς το ταίριασμα των επιμέρους προτύπων που αποτελούν ένα (σύνθετο) πρότυπο. Πράγματι, το FOCUS θεωρεί μόνο μία απλή συνάθροιση των διαφορών των μέτρων στις αντίστοιχες περιοχές. Αυτό το είδος του 1 - 1 ταιριάσματος δεν είναι αρκετά γενικό. Τόσο το παράδειγμα 3 όσο και το παράδειγμα 4 της Ενότητας 3.1 περιγράφουν σενάρια που δεν μπορούν να καλυφθούν από το FOCUS. Στο παράδειγμα 3, λόγου χάρη, γειτονικές περιοχές πρέπει να ληφθούν υπόψη, κάτι που αποκλείει τη χρήση ενός 1 - 1 ταιριάσματος μεταξύ των περιοχών. Από την άλλη, το $PANDA$ είναι πολύ ευέλικτο, δεδομένου ότι ο χρήστης μπορεί εύκολα να ορίσει όλα τα δομικά τμήματα για τη σύγκριση των σύνθετων προτύπων, δηλαδή: (α) πώς θα γίνεται η αποτίμηση της ανομοιότητας μεταξύ απλών προτύπων, (β) πώς θα γίνεται το ταίριασμα των επιμέρους προτύπων και (γ) πώς θα γίνεται η συνάθροιση των τιμών των ταιριασμένων επιμέρους προτύπων προκειμένου να υπολογιστεί η συνολική τιμή της ανομοιότητας.

- Η χρήση του GCR εμποδίζει επίσης την εφαρμογή του FOCUS σε έναν αριθμό από ενδιαφέροντα πρότυπα, για τα οποία δεν ορίζεται η έννοια του GCR, ή όπου ένας πιο ευέλικτος ορισμός του ταιριάσματος μεταξύ των επιμέρους προτύπων απαιτείται.

- Καθώς το $FOCUS$ ασχολείται με τη σύγκριση συνόλων από δεδομένα, απαιτεί την προσπέλαση των δεδομένων αυτών, ενώ η σύγκριση προτύπων στο $PANDA$ γίνεται εξ' ολοκλήρου στο χώρο των προτύπων. Έτσι, το $PANDA$ είναι πιο αποδοτικό (επειδή αποφεύγονται δαπανηρές προσπελάσεις στα δεδομένα) και πιο γενικό (καθώς μπορεί να χειριστεί περιπτώσεις όπου τα πρωτογενή δεδομένα δεν είναι διαθέσιμα). Επίσης, το $PANDA$ λαμβάνει υπόψη και την ιδιωτικότητα των δεδομένων (δεν υπάρχει ανάγκη για προσπέλαση των πρωτογενών δεδομένων που μπορούν να περιέχουν προσωπικές και ευαίσθητες πληροφορίες).

Από τα παραπάνω προκύπτει ότι το $PANDA$ είναι μία επέκταση το FOCUS καθώς χειρίζεται πρότυπα αυθαίρετης πολυπλοκότητας και επιτρέπει μεγαλύτερη ευελιξία στη διαδικασία αποτίμησης της ανομοιότητας.

Επαγωγικές βάσεις δεδομένων Στις επαγωγικές βάσεις δεδομένων (inductive databases) τόσο τα δεδομένα όσο και τα πρότυπα αποθηκεύονται μαζί, με σκοπό να μπορούν να ανακτηθούν και να χειρισθούν με παρόμοιο τρόπο (βλέπε επίσης την Ενότητα 1.3.1). Τα πρότυπα που μελετώνται είναι κυρίως οι κανόνες που εξάγονται από αυτά τα δεδομένα. Στην ουσία, η ανακάλυψη των κανόνων θεωρείται ως ένα ακόμα, ίσως πιο δαπανηρό, είδος ερωτήματος και αυτό εξηγεί γιατί στο πλαίσιο των επαγωγικών βάσεων δεδομένων έχει προταθεί και υλοποιηθεί ένας μεγάλος αριθμός εξειδικευμένων γλωσσών επερωτήσεων.

Οι περισσότερες από αυτές τις γλώσσες επεκτείνουν τη γλώσσα SQL ώστε να υποστηρίζουν και την εξόρυξη γνώσης. Για παράδειγμα, η DMQL [29] είναι μία γλώσσα επερωτήσεων για διάφορους τύπους κανόνων, όπως κανόνες συσχέτισης, γενικευμένες συσχετίσεις και κανόνες χαρακτηριστικών. Η MINE RULE [48] προτείνει ένα μοντέλο που επιτρέπει ομοιόμορφη περιγραφή του προβλήματος εξόρυξης κανόνων συσχέτισης. Η MSQL [34] είναι μία γλώσσα για τη δημιουργία και την αναζήτηση κανόνων συσχέτισης. Σε όλες αυτές τις προσεγγίσεις, τα επαγωγικά ερωτήματα καθορίζουν ενδιαφέροντα πρότυπα με χρήση είτε συντακτικών περιορισμών είτε περιορισμών συχνότητας (ή και των δύο).

Είναι προφανές ότι, στο πλαίσιο των επαγωγικών βάσεων δεδομένων, η σύγκριση προτύπων είναι περιορισμένη σε συγκεκριμένους τύπους προτύπων, όπως οι κανόνες συσχέτισης. Η έμφαση εδώ δίνεται στην ανάκτηση των προτύπων και όχι στη σύγκρισή τους. Τέλος, η περίπτωση των σύνθετων προτύπων δεν λαμβάνεται καθόλου υπόψη.

Παρακολούθηση Προτύπων Μία σχετική γραμμή έρευνας είναι αυτή της παρακολούθησης προτύπων η οποία στοχεύει στην παρακολούθηση των προτύπων στον άξονα του χρονου και στον εντοπισμό των αλλαγών τους.

Σε αυτή τη κατηγορία ανήκει το πλαίσιο DEMON [26] για εξόρυξη συστηματικά εξελισσόμενων δεδομένων στη διάσταση του χρόνου. Ο όρος "συστηματικά" αναφέρεται στις αλλαγές οι οποίες συμβαίνουν λόγω μαζικών προσθηκών ή των διαγραφών εγγραφών από τα δεδομένα, δηλαδή, σύνολα από εγγραφές που προστίθενται ή διαγράφονται συγχρόνως από την βάση δεδομένων. Αρχικά εντοπίζονται τα τμήματα των δεδομένων που έχουν αλλάξει και στη συνέχεια, επεξεργάζονται με σκοπό την ενημέρωση της βάσης των προτύπων. Στα πλαίσια της εργασίας αυτής περιγράφονται αποδοτικοί αλγόριθμοι συντήρησης μοντέλων προτύπων συχνών στοιχειοσυνόλων και συστάδων. Ωστόσο, το DEMON επικεντρώνεται στην αποδοτική ενημέρωση των μοντέλων (δηλαδή προτύπων) παρά στη σύγκρισή τους.

Με το ίδιο σκεπτικό, το πλαίσιο PAM [7] στοχεύει στην αποδοτική συντήρηση των προτύπων (συγκεκριμένα, μελετά κανόνες συσχέτισης) που εξάγονται από ένα δυναμικό σύνολο δεδομένων. Τα πρότυπα μοντελοποιούνται ως χρονικά εξελισσόμενα αντικείμενα που μπορεί να παρουσιάσουν αλλαγές είτε στη δομή τους είτε στις στατιστικές τους ιδιότητες. Προτείνεται ένας μηχανισμός για τον εντοπισμό των αλλαγών στις στατιστικές ιδιότητες των κανόνων.

Πράγματι, η παρακολούθηση προτύπων βασίζεται στη σύγκριση προτύπων με σκοπό τον εντοπισμό σημαντικών αλλαγών στη διάσταση του χρόνου. Αυτό υποδεικνύει μία ακόμα εφαρμογή του προβλήματος εκτίμησης της ανομοιότητας προτύπων και του πλαισίου $PANDA$, αυτή της παρακολούθησης των αλλαγών στα πρότυπα σε δυναμικά περιβάλλοντα.

Ειδικές λύσεις για συγκεκριμένες περιπτώσεις Αρκετές ειδικές λύσεις για τη σύγκριση συγκεκριμένων τύπων προτύπων υπάρχουν στη βιβλιογραφία. Σε αυτή την ενότητα δεν παρουσιάζουμε αυτές τις προσεγγίσεις, καθώς περιγράφονται με λεπτομέρεια στα αντίστοιχα κεφάλαια της διατριβής, συγκεκριμένα στο Κεφάλαιο 4 για τα συχνά στοιχειοσύνολα, στο Κεφάλαιο 5 για τα δέντρα απόφασης και στο Κεφάλαιο 6 για τις συστάδες.

Ο ενδιαφερόμενος αναγνώστης μπορεί να βρει στο [56] μία αναλυτική περιγραφή των μεθόδων σύγκρισης προτύπων στον τομέα της Εξόρυξης Γνώσης, για συχνά

στοιχειοσύνολα και κανόνες συσχέτισης, για συστάδες και συσταδοποιήσεις και για δέντρα απόφασης. Επίσης, στο [47] παρουσιάζεται μία επισκόπηση της σύγκρισης συσταδοποιήσεων που έχουν προκύψει από το ίδιο σύνολο δεδομένων είτε με χρήση διαφορετικών αλγορίθμων είτε με χρήση διαφορετικών παραμέτρων.

3.7 Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάσαμε το $PANDA$ ένα γενικό και ευέλικτο πλαίσιο για αποτίμηση της ανομοιότητας μεταξύ προτύπων αυθαίρετης πολυπλοκότητας. Το $PANDA$ είναι *γενικό* καθώς υποστηρίζει πρότυπα αυθαίρετης πολυπλοκότητας και *ευέλικτο* καθώς η αποτίμηση της ανομοιότητας μπορεί εύκολα να προσαρμοσθεί στις συγκεκριμένες απαιτήσεις του χρήστη ή της εφαρμογής. Στο $PANDA$ τα πρότυπα μοντελοποιούνται ως οντότητες που αποτελούνται από δύο τμήματα: τη *δομική συνιστώσα* η οποία προσδιορίζει τις ενδιαφέρουσες περιοχές στον χώρο των γνωρισμάτων, π.χ., το *head* και το *body* ενός κανόνα συσχέτισης, και την *ποσοτική συνιστώσα* η οποία περιγράφει πώς τα πρότυπα σχετίζονται με τα πρωτογενή δεδομένα από τα οποία έχουν εξαχθεί, π.χ., η υποστήριξη και η εμπιστοσύνη στην περίπτωση ενός κανόνα συσχέτισης. Όταν συγκρίνουμε δύο απλά πρότυπα, η ανομοιότητα των δομικών τους συνιστωσών (*δομική ανομοιότητα*) και η ανομοιότητα των ποσοτικών τους συνιστωσών (*ποσοτική ανομοιότητα*) συνδυάζονται (μέσω μίας *συνάρτησης συνδυασμού*) και έτσι προκύπτει η συνολική τιμή της ανομοιότητας. Το πρόβλημα της σύγκρισης σύνθετων προτύπων αναλύεται στο πρόβλημα της σύγκρισης των αντιστοίχων συνόλων (ή λιστών, πινάκων κτλπ.) των επιμέρους (απλών) προτύπων. Έτσι, πρώτα ταιριάζονται τα επιμέρους πρότυπα (χρησιμοποιώντας έναν συγκεκριμένο *τύπο ταιριάσματος*) και στη συνέχεια, οι τιμές που υπολογίζονται συναθροίζονται (μέσω κάποιας *συνάρτησης συνάθροισης*) προκειμένου να υπολογιστεί η συνολική τιμή ανομοιότητας. Αυτός ο επαναλαμβανόμενος ορισμός της ανομοιότητας επιτρέπει στο $PANDA$ να χειρίζεται πρότυπα αυθαίρετης πολυπλοκότητας.

Ο σκοπός του πλαισίου $PANDA$ είναι να παρέχει στον τελικό χρήστη ένα ισχυρό πλαίσιο για εκτίμηση της ανομοιότητας, ικανό να χειρίζεται συγκεκριμένες απαιτήσεις του χρήστη ή της εφαρμογής. Σε αυτή την εργασία, δεν αντιμετωπίζουμε το πρόβλημα της εύρεσης του καλύτερου μέτρου ανομοιότητας για κάθε δυνατό πρόβλημα, κάτι το οποίο θα ήταν εξάλλου ουτοπικό. Αυτό που παρέχουμε είναι ένας μηχανισμός για τη σύγκριση προτύπων ο οποίος βασίζεται στην ιδέα των απλών/σύνθετων προτύπων. Ο τελικός χρήστης μπορεί να ρυθμίσει τα δομικά τμήματα του $PANDA$ σύμφωνα με τις πραγματικές του ανάγκες και να πειραματιστεί με τα αποτελέσματα έτσι ώστε να αποφασίσει ποιες ρυθμίσεις συνθέτουν το πιο κατάλληλο μέτρο ανομοιότητας για την εκάστοτε εφαρμογή.

Μία πρώιμη έκδοση αυτής της δουλειάς έχει δημοσιευτεί [8], ενώ μία εκτενής έκδοση έχει υποβληθεί [9].

3.8 Ανοιχτά θέματα

Το $PANDA$ είναι ένα πλήρως αρθρωτό πλαίσιο και συνεπώς επιδέχεται διάφορες βελτιώσεις/προεκτάσεις. Μία άμεση επέκταση είναι ο εμπλουτισμός των διαφορετικών δομικών τμημάτων του $PANDA$ (π.χ., dis_{struct}, $Comb$, $Aggr$) με νέες συναρτήσεις έτσι ώστε ο τελικός χρήστης να έχει περισσότερες δυνατότητες επι-

λογής. Στην ίδια κατεύθυνση είναι και ο εμπλουτισμός του $PANDA$ με νέους τύπους προτύπων, όπως τα ακολουθιακά πρότυπα, προκειμένου να διευρυνθεί το σύνολο των υποστηριζόμενων τύπων προτύπων.

Όπως έχουμε ήδη αναφέρει, ανάλογα με το πως αρχικοποιούμε τα διάφορα δομικά τμήματα του $PANDA$ προκύπτουν διαφορετικές συναρτήσεις ανομοιότητας. Η κρίσιμη ερώτηση είναι ποιες από αυτές τις ρυθμίσεις είναι κατάλληλες για το εκάστοτε πρόβλημα. Ήδη υπάρχουσες προσεγγίσεις που έχουν προταθεί στη βιβλιογραφία θα μπορούσαν να ενσωματωθούν στο $PANDA$ προσφέροντας στον τελικό χρήστη έτοιμες, ολοκληρωμένες λύσεις.

Στην παρούσα φάση, το $PANDA$ πραγματεύεται την εκτίμηση της ανομοιότητας μεταξύ προτύπων του ίδιου τύπου. Θα ήταν ενδιαφέρον επίσης να διερευνηθεί και το πρόβλημα της αποτίμησης της ανομοιότητας μεταξύ προτύπων διαφορετικών τύπων. Μία προφανής αντιμετώπιση θα ήταν η μετατροπή του ενός τύπου προτύπων στον άλλο (π.χ., η μετατροπή ενός δέντρου απόφασης σε ένα σύνολο από κανόνες) και η εφαρμογή εν συνεχεία του $PANDA$ πλαισίου. Ωστόσο, αυτή η λύση έχει αρκετά μειονεκτήματα καθώς μία τέτοια μετατροπή δεν είναι πάντα εύκολη. Μία εναλλακτική λύση θα ήταν να οριστεί κάποιος τελεστής ανομοιότητας σε κάποιο υπερ-τύπο προτύπων, του οποίου αποτελούν υπο-τύπους τα προς σύγκριση πρότυπα. Πιο αποδοτικές λύσεις θα πρέπει ωστόσο να διερευνηθούν.

Κεφάλαιο 4

Σύγκριση Συνόλων Δεδομένων μέσω Συχνών Στοιχειοσυνόλων: Επίδραση των Παραμέτρων Εξόρυξης Γνώσης

Στο κεφαλαίο αυτό διερευνούμε το θέμα του κατά πόσο η ανομοιότητα στο χώρο των προτύπων μπορεί να χρησιμοποιηθεί ως μέτρο ανομοιότητας στο χώρο των πρωτογενών δεδομένων. Πιο συγκεκριμένα, εστιάζουμε σε πρότυπα τύπου συχνά στοιχειοσύνολα και μελετάμε πως οι παράμετροι που χρησιμοποιούνται για την εξαγωγή τους επηρεάζουν το αποτέλεσμα της σύγκρισης. Εξετάζουμε δύο τέτοιες παραμέτρους, το κατώφλι *minSupport* και την *αναπαράσταση του πλέγματος* (συχνά στοιχειοσύνολα (frequent itemsets - FI), κλειστά συχνά στοιχειοσύνολα (closed frequent itemsets - CFI), μέγιστα συχνά στοιχειοσύνολα (maximal frequent itemsets - MFI)).

Το κεφάλαιο έχει οργανωθεί ως ακολούθως: Στην Ενότητα 4.1 παρουσιάζουμε μία εισαγωγή στο πρόβλημα. Στην Ενότητα 4.2, περιγράφουμε τις βασικές έννοιες του προβλήματος της εξόρυξης συχνών στοιχειοσυνόλων (frequent itemsets mining -FIM) που απαιτούνται για την κατανόηση του τρέχοντος κεφαλαίου. Στην Ενότητα 4.3, συζητάμε τις σχετικές εργασίες και εισάγουμε ένα γενικό τύπο με βάση τον οποίο μπορούν να περιγραφούν τα διάφορα μέτρα που έχουν προταθεί στη βιβλιογραφία. Στην Ενότητα 4.4, παρουσιάζουμε τις παραμέτρους του ΦΙΜ προβλήματος που επηρεάζουν τη σύγκριση και συγκεκριμένα, το κατώφλι *minSupport* και το επίπεδο συμπίεσης του πλέγματος (FI, CFI, MFI). Στην Ενότητα 4.5, επαληθεύουμε τα θεωρητικά αποτελέσματα ελέγχοντας πειραματικά την επιδραση των διαφόρων παραμέτρων στο αποτέλεσμα της σύγκρισης. Στην Ενότητα 4.6, ανακεφαλαιώνουμε την εργασία μας σε αυτό το κεφάλαιο, ενώ στην Ενότητα 4.7 παρουσιάζουμε τα διάφορα ανοιχτά θέματα.

Λέξεις κλειδιά συχνά στοιχειοσύνολα, μέτρα ομοιότητας, εξάρτηση από παραμέτρους εξόρυξης γνώσης, επιρροή του ελάχιστου κατωφλίου $minSupport$, επιρροή του επιπέδου συμπίεσης του πλέγματος.

4.1 Εισαγωγή

Η εύρεση αλλαγών μεταξύ συνόλων δεδομένων αποτελεί ένα σημαντικό πρόβλημα στις μέρες μας καθώς τα δεδομένα είναι κυρίως δυναμικά και επίσης προέρχονται από διαφορετικές πηγές. Μία συνηθισμένη τεχνική για τη σύγκριση συνόλων δεδομένων είναι να χρησιμοποιήσει κανείς τα σύνολα προτύπων που εξάγονται από τα δεδομένα αυτά μέσω τεχνικών Εξόρυξης Γνώσης. Το κίνητρο πίσω από αυτές τις τεχνικές είναι ότι, σε κάποιο βαθμό, τα πρότυπα συμπυκνώνουν την πληροφορία που υπάρχει στα πρωτογενή δεδομένα. Για παράδειγμα, στην [25], οι συγγραφείς υπολογίζουν την ανομοιότητα μεταξύ δύο συνόλων δεδομένων με βάση τα μοντέλα Εξόρυξης Γνώσης που εξάγονται από τα δεδομένα αυτά (συγκεκριμένα, συχνά στοιχειοσύνολα, τα δέντρα απόφασης και τις συστάδες). Η σύγκριση αυτή χρησιμοποιείται στη συνέχεια στην [26] για την Εξόρυξη Γνώσης από δεδομένα που εξελίσονται με συστηματικό τρόπο, όπως οι αγοραστικές συνήθειες των πελατών ενός σούπερμαρκετ. Στην [66], οι συγγραφείς υπολογίζουν την ανομοιότητα μεταξύ κατανεμημένων συνόλων δεδομένων(π.χ., τη διαφορά μεταξύ των διαφορετικών υποκαταστημάτων ενός σουπερμαρκετ) χρησιμοποιώντας τα αντίστοιχα σύνολα από συχνά στοιχειοσύνολα. Η ίδια λογική ακολουθείται από τους συγγραφείς στην [44].

Διαισθητικά, είναι λογικό να χρησιμοποιεί κανείς την ανομοιότητα στο χώρο των προτύπων για να υπολογίσει την ανομοιότητα στο χώρο των πρωτογενών δεδομένων. Πράγματι, τα πρότυπα διατηρούν μέρος της πληροφορίας που υπάρχει στα πραγματικά δεδομένα (ας θυμηθούμε ότι εξ΄ ορισμού τα πρότυπα αποτελούν συμπαγείς και πλούσιες σε σημασιολογία αναπαραστάσεις των πρωτογενών δεδομένων), ωστόσο το ποσοστό της πληροφορίας που διατηρείται εξαρτάται από τις παραμέτρους της εξόρυξης. Συνεπώς, όταν συγκρίνουμε δύο σύνολα δεδομένων με βάση τα αντίστοιχα σύνολα προτύπων, η υπολογιζόμενη ανομοιότητα είναι υποκειμενική, όσον αφορά στις παραμέτρους της εξόρυξης που χρησιμοποιήσαμε για την εξαγωγή των προτύπων. Το πρόβλημα αυτό σχετίζεται με την Πρόκληση 3 της Ενότητας 1.4.

Στο κεφάλαιο αυτό, εξερευνούμε το προαναφερθέν πρόβλημα για μία πολύ δημοφιλή κατηγορία προτύπων Εξόρυξης Γνώσης, τα συχνά στοιχειοσύνολα (FI) και τις παραλλαγές τους κλειστά συχνά στοιχειοσύνολα (CFI) και μέγιστα κλειστά στοιχειοσύνολα (MFI). Το πρόβλημα της εξόρυξης συχνών στοιχειοσυνόλων έχει μελετηθεί εκτενώς στον τομέα της Εξόρυξης Γνώσης λόγω των πολλαπλών εφαρμογών του (π.χ., κανόνες συσχέτισης, ακολουθιακά πρότυπα και επεισόδια [82]). Πολλές από τις προτεινόμενες εργασίες, εστιάζουν κυρίως στην ανάπτυξη αποδοτικών αλγορίθμων για την επίλυση του προβλήματος (βλέπε [31], [28] για μία επισκόπηση της περιοχής). Επιπλέον, έχουν αναπτυχθεί αλγόριθμοι για την ανακάλυψη πιο συμπαγών αναπαραστάσεων του πλέγματος των συχνών στοιχειοσυνόλων (FI) όπως τα κλειστά συχνά στοιχειοσύνολα (CFI) και τα μέγιστα κλειστά στοιχειοσύνολα (MFI), π.χ., [27], [85].

Για να συνοψίσουμε, στο κεφάλαιο αυτό εξερευνούμε κατά πόσο η ανομοιότητα μεταξύ δύο συνόλων από συχνά στοιχειοσύνολα επηρεάζεται από τις παραμέτρους που χρησιμοποιούνται για την εξόρυξή τους. Η πρώτη παράμετρος που εξετά-

ζουμε είναι το κατώφλι ελάχιστης υποστήριξης $minSupport$ το οποίο περιορίζει το πλέγμα των στοιχειοσυνόλων αφαιρώντας εκείνα τα στοιχειοσύνολα που έχουν υποστήριξη μικρότερη από το δοθέν κατώφλι. Η δεύτερη παράμετρος είναι το *επίπεδο συμπίεσης του πλέγματος* (FI, CFI ή MFI), το οποίο περιορίζει το πλέγμα των συχνών στοιχειοσυνόλων αφαιρώντας τα πλεονάζοντα στοιχειοσύνολα με βάση είτε τη δομική συνίστώσα τους (όπως στα σύνολα MFI) είτε τη δομική και την ποσοτική συνιστώσα (όπως στα σύνολα CFI). Συνεπώς, και οι δύο παράμετροι περιορίζουν το πλέγμα των συχνών στοιχειοσυνόλων με βάση είτε τη δομική, είτε την ποσοτική συνιστώσα ή με βάση και τις δύο συνιστώσες (ποσοτική και δομική).

Η ανάλυσή μας δείχνει πως η χρήση των συνόλων συχνών στοιχειοσυνόλων για τη σύγκριση συνόλων δεδομένων δεν είναι τόσο προφανής όσο παρουσιάζεται από τις σχετικές εργασίες, ένα αποτέλεσμα που επαληθεύεται και μέσα από την πειραματική μας μελέτη και που ανοίγει θέματα για περαιτέρω έρευνα στον τομέα της Εξόρυξης Γνώσης. Η συνεισφορά αυτού του κεφαλαίου συνοψίζεται παρακάτω:

- Παρουσιάζουμε μία θεωρητική ανάλυση που δείχνει την εξάρτηση του αποτελέσματος της σύγκρισης στο χώρο των πρωτύπων από τις παραμέτρους του προβλήματος της εξόρυξης συχνών στοιχειοσυνόλων. Όσον αφορά στην πρώτη παράμετρο, το ελάχιστο κατώφλι υποστήριξης $minSupport$, η ανάλυσή μας δείχνει πως όσο μεγαλύτερο είναι αυτό το κατώφλι τόσο μεγαλύτερη είναι η ανομοιότητα που υπολογίζεται στο χώρο των προτύπων. Όσον αφορά στη δεύτερη παράμετρο, το επίπεδο συμπίεσης του πλέγματος των συχνών στοιχειοσυνόλων (FI, CFI ή MFI), η ανάλυσή μας δείχνει πως όσο πιο συμπιεσμένη είναι η αναπαράσταση του πλέγματος τόσο πιο μεγάλη είναι η ανομοιότητα που υπολογίζεται στο χώρο των προτύπων.

- Περιγράφουμε τα διάφορα μέτρα ανομοιότητας που έχουν προταθεί μέχρι στιγμής [25, 44, 66] μέσω ενός ενιαίου σχήματος ανομοιότητας και επαληθεύουμε πειραματικά τα παραπάνω θεωρητικά αποτελέσματα. Τα αποτελέσματα δείχνουν πως το να συγκρίνουμε σύνολα δεδομένων μέσω των εξαγόμενων συνόλων από πρότυπα δεν είναι τόσο προφανές όσο υποστηρίζουν οι σχετικές εργασίες και θα πρέπει να γίνεται κάτω από συγκεκριμένους περιορισμούς (π.χ., παράμετροι του FIM προβλήματος).

4.2 Βασικές έννοιες

Αρχικά παρουσιάζουμε κάποιες βασικές έννοιες του προβλήματος της εξόρυξης συχνών στοιχειοσυνόλων (FIM) [4]. Έστω I είναι ένα πεπερασμένο σύνολο από διακριτά στοιχεία και D είναι μία βάση δεδομένων συναλλαγών, όπου κάθε συναλλαγή T περιέχει ένα σύνολο στοιχείων, $T \subseteq I$. Ένα *στοιχειοσύνολο* (itemset) X είναι ένα μη κενό λεξικογραφικά διατεταγμένο σύνολο από στοιχεία, $X \subseteq I$. Αν το X περιέχει k στοιχεία, καλείται *k-στοιχειοσύνολο*. Η συχνότητα του X στη D, ισούται με το πλήθος των συναλλαγών της D που περιέχουν το X, δηλαδή, $fr_D(X) = |\{T \in D : X \subseteq T\}|$. Το ποσοστό των συναλλαγών της D που περιέχει το X, καλείται *υποστήριξη* (support) του X στη D· $supp_D(X) = \frac{fr_D(X)}{|D|}$. Ένα *συχνό στοιχειοσύνολο* (frequent itemset) είναι ένα στοιχειοσύνολο με υποστήριξη μεγαλύτερη ή ίση ενός κατωφλίου ελάχιστης υποστήριξης σ που ορίζεται από το χρήστη και καλείται *μινΣυππορτ*, δηλαδή, $supp_D(X) \geq \sigma$.

Το σύνολο των συχνών στοιχειοσυνόλων που εξάγονται από το D με συγκεκριμένο $minSupport$ κατώφλι σ ορίζεται ως εξής:

$$F_\sigma(D) = \{X \subseteq I \mid supp_D(X) \geq \sigma\} \tag{4.1}$$

Το σύνολο των συχνών στοιχειοσυνόλων φτιάχνει το πλέγμα των συχνών στοιχειοσυνόλων στο οποίο ισχύει η λεγόμενη *Apriori* ιδιότητα : ένα στοιχειοσύνολο είναι συχνό αν όλα τα υποσύνολά του είναι επίσης συχνά. Η *apriori* ιδιότητα μας επιτρέπει να απαριθμήσουμε όλα τα συχνά στοιχειοσύνολα χρησιμοποιώντας πιο συμπαγείς μορφές αναπαράστασης όπως τα κλειστά συχνά στοιχειοσύνολα (CFI) και τα μέγιστα συχνά στοιχειοσύνολα (MFI).

Ένα συχνό στοιχειοσύνολο X καλείται *συχνό* αν δεν υπάρχει κανένα συχνό υπερσύνολο $Y \supset X$ με $supp_D(X) = supp_D(Y)$. Το σύνολο των κλειστών συχνών στοιχειοσυνόλων που εξάγεται από το D με συγκεκριμένο $minSupport$ κατώφλι σ ορίζεται ως εξής:

$$C_\sigma(D) = \{X \in F_\sigma(D) : Y \supset X \Rightarrow supp_D(X) > supp_D(Y), Y \in F_\sigma(D)\} \tag{4.2}$$

Το $C_\sigma(D)$ είναι υποσύνολο του $F_\sigma(D)$ καθώς κάθε κλειστό συχνό στοιχειοσύνολο είναι συχνό.

$$C_\sigma(D) = F_\sigma(D) - \{X \in F_\sigma(D) : Y \supset X \Rightarrow supp_D(X) = supp_D(Y), Y \in F_\sigma(D)\} \tag{4.3}$$

Εξ ορισμού, το $C_\sigma(D)$ αποτελεί μία *ακριβή* (lossless) αναπαράσταση του συνόλου $F_\sigma(D)$ καθώς τόσο η δομική όσο και η ποσοτική συνιστώσα των στοιχειοσυνόλων του πλέγματος μπορούν να παραχθούν από το σύνολο των CFI [85].

Από την άλλη, ένα συχνό στοιχειοσύνολο καλείται *μέγιστο* (maximal) αν δεν αποτελεί υποσύνολο κανενός άλλου συχνού στοιχειοσυνόλου. Το σύνολο των μέγιστων συχνών στοιχειοσυνόλων που εξάγεται από το D με συγκεκριμένο $minSupport$ κατώφλι σ ορίζεται ως εξής:

$$M_\sigma(D) = \{X \in F_\sigma(D) : Y \subset X \Rightarrow Y \notin F_\sigma(D), Y \subseteq I\} \tag{4.4}$$

Το $M_\sigma(D)$ είναι επίσης υποσύνολο του $F_\sigma(D)$ καθώς κάθε μέγιστο συχνό στοιχειοσύνολο είναι συχνό.

$$M_\sigma(D) = F_\sigma(D) - \{X \in F_\sigma(D) : Y \supset X \Rightarrow Y \in F_\sigma(D)\} \tag{4.5}$$

Το $M_\sigma(D)$ είναι υπερσύνολο του $C_\sigma(D)$:

$$M_\sigma(D) = C_\sigma(D) - \{X \in C_\sigma(D) : Y \supset X \Rightarrow Y \in C_\sigma(D)\} \tag{4.6}$$

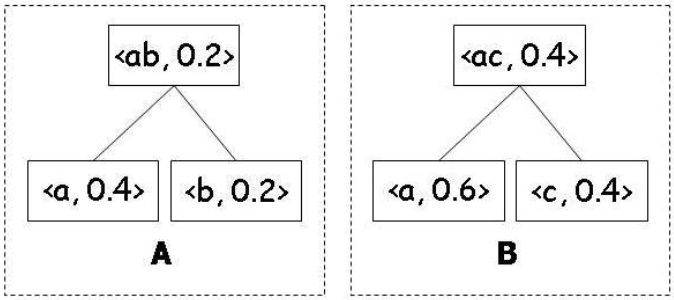
Σε αντίθεση όμως με το σύνολο $C_\sigma(D)$, το σύνολο $M_\sigma(D)$ αποτελεί μία χαλαρή (lossy) αναπαράσταση του $F_\sigma(D)$ καθώς μόνο η δομική συνιστώσα των στοιχειοσυνόλων του πλέγματος (δηλαδή τα ίδια τα στοιχειοσύνολα) μπορεί να εξαχθεί από το σύνολο MFI ενώ η ποσοτική συνιστώσα των (δηλαδή η υποστήριξη των στοιχειοσυνόλων) έχει χαθεί [85].

Από τις Εξίσωση 4.3, 4.5 και 4.6 συνεπάγεται ότι:

$$M_\sigma(D) \subseteq C_\sigma(D) \subseteq F_\sigma(D) \tag{4.7}$$

Στην πράξη, το σύνολο CFI μπορεί να είναι τάξεις μεγέθους μικρότερο από το σύνολο FI, και το σύνολο MFI μπορεί να είναι τάξεις μεγέθους μικρότερο από το σύνολο CFI [85]· βλέπε επίσης το Σχήμα 4.3.

Για διευκόλυνση στην κατανόηση, ας θεωρήσουμε τα δύο σύνολα από συχνά στοιχειοσύνολα A, B που παρουσιάζονται στο Σχήμα 4.1. Έστω ότι τα σύνολα A, B έχουν προκύψει εφαρμόζοντας το ίδιο $minSupport$ κατώφλι σ στα αρχικά σύνολα δεδομένων D και E, αντίστοιχα. Κάθε στοιχειοσύνολο περιγράφεται ως ένα ζεύγος <*δομή, μέτρο*> που συμβολίζουν τα στοιχεία που απαρτίζουν το στοιχειοσύνολο (*δομική συνιστώσα*) και την υποστήριξη του στοιχειοσυνόλου (*ποσοτική συνιστώσα*).



Σχήμα 4.1: Τα πλέγματα συχνών στοιχειοσυνόλων: A (αριστερά), B (δεξιά)

Η ερώτηση είναι πόσο ανόμοια είναι τα δύο σύνολα A και B. Υπάρχουν αρκετές περιπτώσεις όπου τα δύο σύνολα μπορεί να διαφέρουν: Για παράδειγμα, ένα στοιχειοσύνολο μπορεί να εμφανίζεται και στα δύο σύνολα αλλά με διαφορετική υποστήριξη, όπως το στοιχειοσύνολο $< a >$ στο Σχήμα 4.1. Ή, ένα στοιχειοσύνολο μπορεί να εμφανίζεται μόνο σε ένα από τα δύο σύνολα, όπυ το στοιχειοσύνολο $< b >$ που εμφανίζεται στο σύνολο A αλλά όχι στο B. Στην περίπτωση αυτή, δύο πράγματα ενδέχεται να έχουν συμβεί: είτε το $< b >$ δεν υπάρχει πραγματικά στο αντίστοιχο σύνολο δεδομένων E ή το $< b >$ έχει απομακρυνθεί λόγω χαμηλής υποστήριξης (μικρότερης από το $minSupport$ κατώφλι σ).

Επειδή τα σύνολα A, B εξαρτώνται από τις παραμέτρους εξόρυξης γνώσης και συγκεκριμένα από το $minSupport$ κατώφλι σ που χρησιμοποιήθηκε για την εξαγωγή τους και από το επίπεδο συμπίεσης του πλέγματος (FI, CFI ή MFI), υποστηρίζουμε ότι και η ανομοιότητα τους εξαρτάται από τις παραμέτρους αυτές. Επίσης, επειδή η ανομοιότητα στο χώρο των προτύπων χρησιμοποιείται ως μέτρο της ανομοιότητας στο χώρο των πρωτογενών δεδομένων, υποστηρίζουμε ότι οι προαναφερθείσες παράμετροι επηρεάζουν και την αντιστοιχία αυτή.

Στον Πίνακα 4.1 παρουσιάζονται τα σύμβολα που χρησιμοποιούνται στο κεφάλαιο αυτό.

4.3 Σύγκριση πλεγμάτων συχνών στοιχειοσυνόλων

Στη σχετική βιβλιογράφία ([25], [44], [66]), η σύγκριση μεταξύ συνόλων από στοιχειοσύνολα χρησιμοποιείται για τη σύγκριση συνόλων από δεδομένα. Τόσο η δομική όσο και η ποσοτική συνιστώσα των στοιχειοσυνόλων αξιοποιούνται στα πλαίσια αυτής της σύγκρισης. Στις επόμενες παραγράφους περιγράφουμε τις προσεγγίσεις που έχουν προταθεί - δεν αναφερόμαστε στην [81], η οποία δεν λαμβάνει υπόψη την ποσοτική συνιστώσα των προτύπων.

Σύμβολο	Περιγραφή
D	ένα σύνολο δεδομένων
X	ένα στοιχειοσύνολο
$supp_D(X)$	η υποστήριξη του X στο D
σ	το κατώφλι ελάχιστης υποστήριξης, $minSupport$
$F_\sigma(D)$	το σύνολο των συχνών στοιχειοσυνόλων του D με βάση το σ
$C_\sigma(D)$	το σύνολο των κλειστών συχνών στοιχειοσυνόλων του D με βάση το σ
$M_\sigma(D)$	το σύνολο των μέγιστων συχνών στοιχειοσυνόλων του D με βάση το σ
$dis(A,B)$	η ανομοιότητα μεταξύ δύο συνόλων από στοιχειοσύνολα A, B

Πίνακας 4.1: Λίστα συμβόλων για το Κεφάλαιο 4

4.3.1 Η προσέγγιση των $Parthasarathy - Ogihara$

Οι Parthasarathy και Ogihara [66] παρουσιάζουν μία μέθοδο για την αποτίμηση της ανομοιότητας μετάξύ δύο συνόλων δεδομένων D και E, η οποία χρησιμοποιεί τα σύνολα στοιχειοσυνόλων που εξάγονται από αυτά (A και B, αντίστοιχα). Η μετρική τους ορίζεται ως εξής:

$$dis(A,B) = 1 - \frac{\sum_{X \in A \cap B} \max\{0, 1 - \theta * |supp_D(X) - supp_E(X)|\}}{|A \cup B|} \quad (4.8)$$

Στην παραπάνω εξίσωση, η παράμετρος θ είναι μία παράμετρος κλιμάκωσης που ορίζεται από το χρήστη και ανταναχλά πόσο σημαντικές είναι για το χρήστη οι διακυμάνσεις στις τιμές της υποστήριξης των στοιχειοσυνόλων. Για $\theta = 0$, η ποσοτική συνιστώσα των στοιχειοσυνόλων δεν λαμβάνεται κάθόλου υπόψη, ενώ για $\theta = 1$, η ποσοτική συνιστώσα είναι το ίδιο σημαντική με τη δομική συνιστώσα.

Αυτό το μέτρο δουλεύει με στοιχειοσύνολα με ίδια δομική συνιστώσα, δηλαδή με αυτά που εμφανίζονται στο σύνολο $A \cap B$. Τα στοιχειοσύνολα που μόνο μερικώς μοιάζουν, όπως τα $< ab >$ και $< ac >$ στο παράδειγμά μας, δεν προσδίδουν μεγαλύτερη ομοιότητα στα δύο σύνολα.

Για το παράδειγμά μας, ισχύει ότι $A \cap B = \{< a >\}$. Αν υποθέσουμε ότι $\theta = 1$, προκύπτει ότι $dis(A,B) = 0.84$ με βάση την Εξίσωση 4.8.

4.3.2 Η προσέγγιση του $FOCUS$

Στην εργασία [25], οι Ganti et al προτείνουν το πλαίσιο FOCUS για την αποτίμηση της ανομοιότητας μέταξύ δύο συνόλων δεδομένων D και E, το οποίο στηρίζεται στην ανομοιότητα των σύνολων από στοιχειοσύνολα (A και B, αντίστοιχα) που εξάγονται από τα δεδομένα αυτά. Η ανομοιότητα ορίζεται ως το ποσό της εργασίας που απατείται για την μετατροπή του ενός συνόλου στο άλλο. Για το σκοπό αυτό, τα A και B σύνολα εκλεπτύνονται στην ένωσή τους $A \cup B$ και η υποστήριξη κάθε στοιχειοσυνόλου που ανήκει στην ένωση υπολογίζεται με βάση τα δύο σύνολα δεδομένων D και E. Στη συνέχεια, η ανομοιότητα υπολογίζεται αθροίζοντας τις διαφορές στην υποστήριξη όλων των στοιχειοσυνόλων που ανήκουν στην ένωση $A \cup B$:

$$dis(A,B) = \frac{\sum_{X \in A \cup B} |supp_D(X) - supp_E(X)|}{\sum_{X \in A} supp_D(X) + \sum_{X \in B} supp_E(X)} \quad (4.9)$$

Το πλαίσιο FOCUS υπολογίζει την ανομοιότητα μεταξύ δύο συνόλων από στοιχειοσύνολα με βάση τα στοιχειοσύνολα που ανήκουν στην ένωση των δύο

συνόλων. Η έννοια της μερικής ομοιότητας δεν υπάρχει στο FOCUS. Στην πραγματικότητα, το FOCUS προσπαθεί να βρει και στα δύο σύνολα στοιχειοσύνολα με την ίδια δομή. Έτσι, αν ένα στοιχειοσύνολο εμφανίζεται τόσο στο σύνολο A όσο και στο σύνολο B, η διαφορά στην τιμή της υποστήριξής τους υπολογίζεται και προστίθεται στο τελικό σκορ. Αν όμως, ένα στοιχειοσύνολο X εμφανίζεται μόνο στο σύνολο A με $supp_D(X)$ αλλά δεν εμφανίζεται στο σύνολο B, τότε το FOCUS θέτει ένα επερώτημα στο αντίστοιχο σύνολο δεδομένων του B, δηλαδή στο E, και το $supp_E(X)$ υπολογίζεται.

Επειδή η επερώτηση των αρχικών συνόλων δεδομένων είναι μία ακριβή διαδικασία, οι συγγραφείς παρέχουν ένα άνω όριο ανομοιότητας για την περίπτωση των συχνών στοιχειοσυνόλων, το οποίο δεν απαιτεί την επερώτηση των αρχικών συνόλων δεδομένων αλλά περιορίζεται στα σύνολα στοιχειοσυνόλων. Στην περίπτωση αυτή, αν ένα στοιχειοσύνολο X δεν εμφανίζεται στο B, θεωρείται ότι εμφανίζεται αλλά με μηδενική υποστήριξη, δηλαδή, $supp_E(X) = 0$. Στη συνέχεια, θεωρούμε ως μέτρο ανομοιότητας του FOCUS αυτό το άνω όριο, καθώς σκοπός της εργασίας μας είναι να δούμε κατά πόσο τα πρότυπα (δηλαδή, τα σύνολα από συχνά στοιχειοσύνολα) διατηρούν τυχόν χαρακτηριστικά ομοιότητας που υπάρχουν στα πρωτογενή σύνολα δεδομένων.

Για το παράδειγμά μας, ισχύει ότι $A \cup B = \{< a >, < b >, < c >, < ab >, < ac >\}$. Με βάση την Εξίσωση 4.9, προκύπτει ότι $dis(A, B) = 0.64$.

4.3.3 Η προσέγγιση των $Li - Ogihara - Zhou$

Οι Li et al [44] προτείνουν ένα μέτρο ομοιότητας μεταξύ δύο συνόλων από δεδομένα με βάση τα σύνολα από μέγιστα συχνά στοιχειοσύνολα (MFI) που εξάγονται από τα δεδομένα αυτά. Το μέτρο τους ορίζεται ως εξής: Έστω $A = \{X_i, supp_D(X_i)\}$ και $B = \{Y_j, supp_E(Y_j)\}$ όπου X_i, Y_j είναι τα MFI των D και E αντίστοιχα. Η ανομοιότητα μεταξύ των συνόλων A και B ορίζεται ως εξής:

$$dis(A, B) = 1 - \frac{2I_3}{I_1 + I_2} \tag{4.10}$$

όπου

$$I_1 = \sum_{X_i, X_j \in A} d(X_i, X_j), \quad I_2 = \sum_{Y_i, Y_j \in B} d(Y_i, Y_j), \quad I_3 = \sum_{X \in A, Y \in B} d(X, Y)$$

και

$$d(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} * log(1 + \frac{|X \cap Y|}{|X \cup Y|}) * \min(supp_D(X), supp_E(Y))$$

Το I_3 μπορεί να θεωρηθεί ως μέτρο της " αμοιβαίας πληροφορίας " που υπάρχει μεταξύ των συνόλων A και B, ενώ το κλάσμα $\frac{2}{I_1+I_2}$ χρησιμοποιείται ως παράγοντας κανονικοποίησης.

Το μέτρο αυτό δουλεύει με τη μέση ανομοιότητα μεταξύ ζευγών από MFI των συνόλων A και B. Η έννοια της μερικής ανομοιότητας υποστηρίζεται από την προσέγγιση αυτή, συνεπώς στοιχειοσύνολα που έχουν παρόμοια δομή συγκρίνονται και το αποτέλεσμα της σύγκρισής τους συνεισφέρει στο τελικό σκορ, $dis(A, B)$.

Για το παράδειγμα μας, ισχύει ότι $dis(A, B) = 0.58$ με βάση την Εξίσωση 4.10.

4.3.4 Ένας γενικός τύπος για τις τρεις προτεινόμενες προσεγγίσεις

Και οι τρεις προσεγγίσεις που παρουσιάσαμε εκφράζουν την ανομοιότητα μετάξύ δύο συνόλων από συχνά στοιχειοσύνολα συναθροίζοντας τις αναμοιότητες των επιμέρους στοιχειοσυνόλων (Στον τύπο αυτό δεν θεωρούμε τον παράγοντα κανονικοποίησης.):

$$dis(A,B) = \sum_{X \in A, Y \in B} dis(X,Y) \tag{4.11}$$

όπου $dis(X,Y)$ είναι η ανομοιότητα μεταξύ δύο απλών συχνών στοιχειοσυνόλων, που ορίζεται με βάση τη δομική και την ποσοτική συνιστώσα των στοιχειοσυνόλων, ως ακολούθως:

$$dis(X,Y) = \varphi(dis_{struct}(X,Y), dis_{meas}(X,Y)) \tag{4.12}$$

Η συνάρτηση $dis_{struct}(X,Y)$ υπολογίζει την ανομοιότητα μεταξύ των δομικών συνιστωσών των X, Y, ενώ η συνάρτηση $dis_{meas}(X,Y)$ υπολογίζει την ανομοιότητα μεταξύ των ποσοτικών συνιστωσών (δηλαδή, της υποστήριξης). Τέλος, η συνάρτηση φ συνδυάζει τις επιμέρους διαφορές όσον αφορά στις δομικές και τις ποσοτικές συνιστώσες σε ένα συνολικό σκορ.

Και οι τρεις προσεγγίσεις ([66], [25], [44]) ακολουθούν την ίδια λογική της Εξίσωση 4.11 και διαφοροποιούνται όσον αφορά στον τρόπο υπολογισμού των $dis_{struct}(X,Y)$ και $dis_{meas}(X,Y)$ και όσον αφορά στο πως τα επιμέρους σκορ συνδυάζονται μέσω της συνάρτησης φ.

- Στην περίπτωση της προσέγγισης των Parthasarathy-Ogihara, η Εξίσωση 4.12 μπορεί να γραφτεί ως εξής:

  $$dis(X,Y) = \max\{0, 1 - dis_{struct}(X,Y) - \theta * dis_{meas}(X,Y)\}$$

  όπου

  $$dis_{struct}(X,Y) = \begin{cases} 0 & \text{, αν } X = Y \\ 1 & \text{, ειδάλλως} \end{cases}$$

  $$dis_{meas}(X,Y) = \begin{cases} |supp_D(X) - supp_E(Y)| & \text{, αν } X = Y \\ 0 & \text{, ειδάλλως} \end{cases}$$

- Στην περίπτωση της προσέγγισης του FOCUS, η Εξίσωση 4.12 μπορεί να γραφτεί ως εξής:

  $$dis(X,Y) = (1 - dis_{struct}(X,Y)) * dis_{meas}(X,Y)$$

  όπου

  $$dis_{struct}(X,Y) = \begin{cases} 0 & \text{, αν } X = Y \\ 1 & \text{, ειδάλλως} \end{cases}$$

  $$dis_{meas}(X,Y) = \begin{cases} |supp_D(X) - supp_E(Y)| & \text{, αν } X,Y \in A \cap B \text{ και } X = Y \\ supp_D(X) & \text{, αν } X \in A - B \\ supp_E(Y) & \text{, αν } Y \in B - A \end{cases}$$

- Τέλος, στην περίπτωση της προσέγγισης των Li-Ogihara-Zhu, η Εξίσωση 4.12 μπορεί να γραφτεί ως εξής:

$$dis(X,Y) = dis_{struct}(X,Y) * \log(1 + dis_{struct}(X,Y)) * dis_{meas}(X,Y)$$

όπου

$$dis_{struct}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

$$dis_{meas}(X,Y) = \min\{supp_D(X), supp_E(Y)\}$$

Το γεγονός ότι όλες οι προσεγγίσεις ακολουθούν ένα γενικό τύπο αποδεικνύεται και στα πειράματα (Ενότητα 4.5), όπου, παρόλο που οι τιμές της ανομοιότητας διαφέρουν μεταξύ των διαφορετικών προσεγγίσεων, μία κοινή συμπεριφορά διαφαίνεται.

Θα μπορούσαμε να εκφράσουμε τα διάφορα προτεινόμενα μέτρα με βάση το πλαίσιο $PANDA$ (βλέπε Κεφάλαιο 3): τα στοιχειοσύνολα αντιστοιχούν σε απλά πρότυπα ενώ τα σύνολα στοιχειοσυνόλων αντιστοιχούν σε σύνθετα πρότυπα. Ο τρόπος υπολογισμού της ανομοιότητας μεταξύ δύο απλών προτύπων (δηλαδή, στοιχειοσυνόλων) για κάθε προσέγγιση παρουσιάστηκε ήδη σε αυτή την ενότητα. Για τη σύγκριση σύνθετων προτύπων (δηλαδή, συνόλων στοιχειοσυνόλων) όλες οι προσεγγίσεις χρησιμοποιούν τη συνάρτηση sum ως συνάρτηση συνάθροισης. Όσον αφορά στον τύπο ταιριάσματος, η προσέγγιση των Parthasarathy-Ogihara και του FOCUS ακολουθεί το $1-1$ *ταίριασμα*, ενώ η προσέγγιση των Li-Ogihara-Zhou ακολουθεί το $M-N$ *ταίριασμα* μεταξύ των επιμέρους στοιχειοσυνόλων.

4.4 Επίδραση των παραμέτρων Εξόρυξης Γνώσης στην ομοιότητα

Στις επόμενες παραγράφους, εξερευνούμε πως η ανομοιότητα μεταξύ δύο συνόλων από στοιχειοσύνολα επηρεάζεται από το $minSupport$ κατώφλι σ που χρησιμοποιείται για την εξαγωγή τους καθώς επίσης και από το επίπεδο συμπίεσης του πλέγματος (FI, CFI ή MFI). Ξεκινάμε με ένα σύνολο δεδομένων από το οποίο εξάγουμε τα συχνά στοιχειοσύνολα εφαρμόζοντας διαφορετικές παραμέτρους εξόρυξης και στη συνέχεια συγκρίνουμε τα σύνολα προτύπων που προκύπτουν με το αρχικό σύνολο προτύπων (δηλαδή, αυτό που εξάχθηκε από το αρχικό σύνολο δεδομένων).

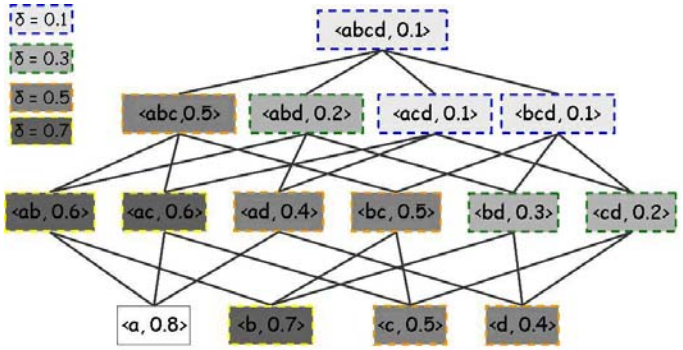
4.4.1 Επίδραση του κατωφλίου $minSupport$ στο αποτέλεσμα της σύγκρισης

Έστω σ, $\sigma+\delta$ $(0 < \sigma, \delta < \sigma+\delta \leq 1)$ είναι δύο τιμές για το $minSupport$ κατώφλι. Έστω επίσης F_σ, $F_{\sigma+\delta}$ είναι τα αντίστοιχα σύνολα από στοιχειοσύνολα που προκύπτουν με βάση τις τιμές αυτές . Το σύνολο $F_\sigma - F_{\sigma+\delta}$ περιέχει όλα τα στοιχειοσύνολα των οποίων η υποστήριξη κυμαίνεται μεταξύ των τιμών σ και $\sigma+\delta$:

$$Z \equiv F_\sigma - F_{\sigma+\delta} = \{X \subseteq I \mid \sigma \leq supp(X) < \sigma+\delta\} \tag{4.13}$$

Στο Σχήμα 4.2, παραθέτουμε ένα παράδειγμα όπου φαίνεται πως το πλέγμα επηρεάζεται από την αύξηση δ του $minSupport$ κατωφλίου. Όπως φαίνεται στην

εικόνα αυτή, καθώς η τιμή του δ αυξάνεται, το εναπομένον πλέγμα μειώνεται. Να σημειώσουμε ωστόσο, ότι δύο τιμές δ, δ' μπορεί να παράγουν το ίδιο πλέγμα ακόμα και αν $\delta < \delta'$, συνεπώς, η συνάρτηση " κλαδέματος " του πλέγματος με βάση το $minSupport$ κατώφλι δεν είναι μονοτονική. Για παράδειγμα, στοΣχήμα 4.2 προκύπτει το ίδιο πλέγμα είτε εφαρμόζοντας $\delta = 0.05$ είτε εφαρμόζοντας $\delta' = 0.1$.



Σχήμα 4.2: Επίδραση της αύξησης του δ στη δομή του πλέγματος ($\sigma = 0.1$)

Παρακάτω, περιγράφουμε πως τα προαναφερόμενα μέτρα επηρεάζονται από την αύξηση δ στην τιμή του $minSupport$ κατωφλίου.

4.4.1.1 Η προσέγγιση των *Parthasarathy-Ogihara* [66]

Λαμβάνοντας υπόψη την Εξίσωση 4.8 και δεδομένου ότι $F_\sigma \cap F_{\sigma+\delta} = F_{\sigma+\delta}$, προκύπτει ότι:

$$dis(F_\sigma, F_{\sigma+\delta}) = 1 - \frac{\sum_{X \in F_\sigma \cap F_{\sigma+\delta}} \max\{0, 1 - \theta * |supp_D(X) - supp_D(X)|\}}{|F_\sigma \cup F_{\sigma+\delta}|}$$
$$= 1 - \frac{\sum_{X \in F_{\sigma+\delta}} \max\{0, 1-0\}}{|F_\sigma|}$$
$$\Rightarrow dis(F_\sigma, F_{\sigma+\delta}) = 1 - \frac{|F_{\sigma+\delta}|}{|F_\sigma|} \tag{4.14}$$

Από την παραπάνω εξίσωση προκύπτει ότι όσο μεγαλύτερη είναι η αύξηση στο $minSupport$ κατώφλι δ, τόσο μικρότερη είναι η τιμή του αριθμητή $|F_{\sigma+\delta}|$ (βλέπε Εξίσωση 4.13) και συνεπώς τόσο μεγαλύτερη είναι η απόσταση μεταξύ των δύο συνόλων.

4.4.1.2 Η προσέγγιση του *FOCUS* [25]

Με βάση τις εξισώσεις Εξίσωση 4.9 και Εξίσωση 4.13, προκύπτει ότι:

$$dis(F_\sigma, F_{\sigma+\delta}) = \frac{\sum_{X \in F_\sigma \cup F_{\sigma+\delta}} |supp_D(X) - supp_D(X)|}{\sum_{X \in F_\sigma} supp_D(X) + \sum_{X \in F_{\sigma+\delta}} supp_D(X)} \tag{4.15}$$
$$= \frac{\sum_{X: \sigma \leq supp_D(X) < \sigma+\delta} supp_D(X)}{2 * \sum_{X \in F_\sigma} supp_D(X) - \sum_{X: \sigma \leq supp_D(X) < \sigma+\delta} supp_D(X)}$$

Επειδή ο συμβολισμός σε αυτή την εξίσωση μπορεί να μπερδέψει τον τελικό αναγνώστη, παρέχουμε εδώ κάποιες περαιτέρω εξηγήσεις. Συγκεκριμένα, ο όρος $\sum_{X \in F_\sigma \cup F_{\sigma+\delta}} |supp_D(X) - supp_D(X)|$ αντιστοιχεί στο άθροισμα των τιμών υποστήριξης για όλα τα στοιχειοσύνολα που εμφανίζονται στη διαφορά των δύο συνόλων $F_\sigma - F_{\sigma+\delta}$. Αν αυτό το σύνολο δεν είναι κενό, ο παραπάνω όρος είναι > 0.

Για λόγους απλότητας, έστω $C = \sum_{X:\sigma \leq supp_D(X) < \sigma+\delta} supp_D(X)$, δηλαδή το C περιλαμβάνει όλα εκείνα τα στοιχειοσύνολα με τιμές υποστήριξης μεταξύ σ και $\sigma + \delta$. Συνεπώς:

$$\Rightarrow dis(F_\sigma, F_{\sigma+\delta}) = \frac{C}{2 * \sum_{X \in F_\sigma} supp_D(X) - C} \tag{4.16}$$

Στην παραπάνω εξίσωση, αν αυξηθεί η τιμή του δ, ο αριθμητής (C) θα αυξηθεί επίσης, ενώ ο παρανομαστής θα μειωθεί (βλέπε Εξίσωση 4.13 επίσης). Συνεπώς, όσο το δ αυξάνεται τόσο η ανομοιότητα αυξάνεται.

4.4.1.3 Η προσέγγιση των *Li-Ogihara-Zhou* [44]

Με βάση την Εξίσωση 4.10 και την Εξίσωση 4.13, προκύπτει ότι:

$$\begin{aligned} I_1 + I_2 &= \sum_{X,Y \in F_\sigma} d(X,Y) + \sum_{X,Y \in F_{\sigma+\delta}} d(X,Y) \\ &= 2 * \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{\substack{X:\sigma \leq supp(X) < \sigma+\delta \\ Y:\sigma \leq supp(Y) < \sigma+\delta}} d(X,Y) \end{aligned}$$

$$I_3 = \sum_{\substack{X \in F_\sigma \\ Y \in F_{\sigma+\delta}}} d(X,Y) = \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{\substack{X:\sigma \leq supp(X) < \sigma+\delta \\ Y:\sigma \leq supp(Y) < \sigma+\delta}} d(X,Y)$$

Για λόγους απλότητας, έστω $G = \sum_{\substack{X:\sigma \leq supp(X) < \sigma+\delta \\ Y:\sigma \leq supp(X) < \sigma+\delta}} d(X,Y)$, δηλαδή το G περιλαμβάνει όλα εκείνα τα στοιχειοσύνολα με τιμές υποστήριξης μεταξύ σ και $\sigma + \delta$. Συνεπώς:

$$\Rightarrow dis(F_\sigma, F_{\sigma+\delta}) = 1 - \frac{2I_3}{I_1 + I_2} = 1 - \frac{2(I_1 - G)}{2I_1 - G} = \frac{G}{2I_1 - G} \tag{4.17}$$

Από την παραπάνω εξίσωση προκύπτει ότι καθώς το δ αυξάνεται, ο αριθμητής (G) αυξάνεται επίσης, ενώ ο παρανομαστής ($2I_1 - G$) μειώνεται (βλέπε Εξίσωση 4.13 επίσης). Συνεπώς, η ανομοιότητα αυξάνεται καθώς το δ αυξάνεται.

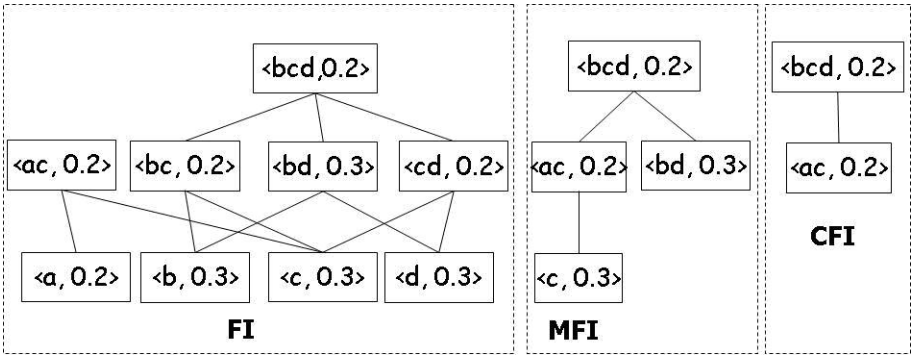
Συνοψίζοντας, οι Εξίσωση 4.14, 4.16 και 4.17 δηλώνουν ότι, για όλες τις προσεγγίσεις, όσο μεγαλύτερη είναι η αύξηση στην τιμή του $minSupport$ κατωφλίου δ, τόσο μεγαλύτερη είναι η υπολογιζόμενη ανομοιότητα, $dis(F_\sigma, F_{\sigma+\delta})$.

4.4.2 Επίδραση του επιπέδου συμπίεσης του πλέγματος στο αποτέλεσμα της σύγκρισης

Έστω $F_\sigma(D)$, $C_\sigma(D)$, $M_\sigma(D)$ είναι τα σύνολα από συχνά, κλειστά συχνά και μέγιστα συχνά στοιχειοσύνολα, αντίστοιχα τα οποία έχουν εξαχθεί από ένα σύνολο

δεδομένων D κάτω από συγκεκριμένο $minSupport$ κατώφλι σ. Στην υποενότητα αυτή, μελετάμε πως η αναπαράσταση των στοιχειοσυνόλων που υιοθετούμε επηρεάζει την υπολογιζόμενη ανομοιότητα.

Στο Σχήμα 4.3 παραθέτουμε ένα παράδειγμα όπου φαίνεται η επίδραση των διαφορετικών αναπαραστάσεων στοιχειοσυνόλων (FI, CFI, MFI) στη δομή του πλέγματος. Αυτές οι εικόνες επιβεβαιώνουν την Εξίσωση 4.7 που λέει πως όσο μεγαλύτερο είναι το επίπεδο συμπίεσης των στοιχειοσυνόλων τόσο πιο μικρό είναι το εναπομένον πλέγμα.



Σχήμα 4.3: Επίδραση των διαφόρων αναπαραστάσεων στοιχειοσυνόλων (FI, CFI, MFI) στο πλέγμα ($\sigma = 0.1$)

Στη συνέχεια περιγράφουμε, για καθένα από τα παρουσιαζόμενα μέτρα ανοιμοιότητας, πως επηρεάζεται από την υιοθετημένη αναπαράσταση στοιχειοσυνόλων (FI, CFI, MFI) για το πλέγμα.

4.4.2.1 Η προσέγγιση των *Parthasarathy-Ogihara* [66]

Από την Εξίσωση 4.8, και επειδή $F_\sigma \cap C_\sigma = C_\sigma, F_\sigma \cap M_\sigma = M_\sigma$, προκύπτει ότι:

$$dis(F_\sigma, C_\sigma) = 1 - \frac{\sum_{X \in F_\sigma \cap C_\sigma} \max\{0, 1 - \theta * |supp_D(X) - supp_D(X)|\}}{|F_\sigma \cup C_\sigma|} = 1 - \frac{\sum_{X \in C_\sigma} \max\{0, 1 - 0\}}{|F_\sigma|} = 1 - \frac{|C_\sigma|}{|F_\sigma|} \quad (4.18)$$

επίσης:

$$dis(F_\sigma, M_\sigma) = 1 - \frac{\sum_{X \in F_\sigma \cap M_\sigma} \max\{0, 1 - \theta * |supp_D(X) - supp_D(X)|\}}{|F_\sigma \cup M_\sigma|} = 1 - \frac{\sum_{X \in M_\sigma} \max\{0, 1 - 0\}}{|F_\sigma|} = 1 - \frac{|M_\sigma|}{|F_\sigma|} \quad (4.19)$$

Από τις δύο παραπάνω εξισώσεις, ισχύει, για την περίπτωση της προσέγγισης των Parthasarathy-Ogihara, ότι:

$$dis(F_\sigma, C_\sigma) \leq dis(F_\sigma, M_\sigma) \quad (4.20)$$

4.4.2.2 Η προσέγγιση του *FOCUS* [25]

Με βάση την Εξίσωση 4.9, και επειδή $F_\sigma \cup C_\sigma = F_\sigma, F_\sigma \cup M_\sigma = F_\sigma$, ισχύει ότι:

$$dis(F_\sigma, C_\sigma) = \frac{\sum_{X \in F_\sigma \cup C_\sigma} |supp_D(X) - supp_D(X)|}{\sum_{X \in F_\sigma} supp_D(X) + \sum_{X \in C_\sigma} supp_D(X)} = \frac{\sum_{X \in F_\sigma - C_\sigma} supp_D(X)}{2 * \sum_{X \in F_\sigma} supp_D(X) - \sum_{X \in F_\sigma - C_\sigma} supp_D(X)} \tag{4.21}$$

επίσης:

$$dis(F_\sigma, M_\sigma) = \frac{\sum_{X \in F_\sigma \cup M_\sigma} |supp_D(X) - supp_D(X)|}{\sum_{X \in F_\sigma} supp_D(X) + \sum_{X \in M_\sigma} supp_D(X)} = \frac{\sum_{X \in F_\sigma - M_\sigma} supp_D(X)}{2 * \sum_{X \in F_\sigma} supp_D(X) - \sum_{X \in F_\sigma - M_\sigma} supp_D(X)} \tag{4.22}$$

ωηερε $F_\sigma - C_\sigma$ ις γιεν βψ Εξίσωση 4.3 ανδ $F_\sigma - M_\sigma$ ις γιεν βψ Εξίσωση 4.5.

Από τις παραπάνω δύο εξισώσεις, ισχύει, για την προσέγγιση του FOCUS, ότι:

$$dis(F_\sigma, C_\sigma) \leq dis(F_\sigma, M_\sigma) \tag{4.23}$$

4.4.2.3 Η προσέγγιση των *Li-Ogihara-Zhou* [44]

Από την Εξίσωση 4.10, προκύπτει ότι:

$$I_1 + I_2 = \sum_{X,Y \in F_\sigma} d(X,Y) + \sum_{X,Y \in C_\sigma} d(X,Y)$$

$$= 2 * \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y) = 2 * I_1 - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$$

$$I_3 = \sum_{\substack{X \in F_\sigma \\ Y \in C_\sigma}} d(X,Y) = \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$$

$$= I_1 - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$$

Για λόγους απλότητας, έστω $K = \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$. Συνεπώς:

$$dis(F_\sigma, C_\sigma) = 1 - \frac{2(I_1 - K)}{2I_1 - K} = \frac{K}{2I_1 - K} \tag{4.24}$$

Με παρόμοιο τρόπο, για την περίπτωση της σύγκρισης $FI - MFI$ ισχύει ότι:

$$I_1 + I_2 = \sum_{X,Y \in F_\sigma} d(X,Y) + \sum_{X,Y \in M_\sigma} d(X,Y)$$

$$= 2 * \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y) = 2 * I_1 - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$$

$$I_3 = \sum_{\substack{X \in F_\sigma \\ Y \in M_\sigma}} d(X,Y) = \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$$

$$= I_1 - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$$

Για λόγους απλότητας, έστω $L = \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$. Συνεπώς:

$$dis(F_\sigma, M_\sigma) = 1 - \frac{2(I_1 - L)}{2I_1 - L} = \frac{L}{2I_1 - L} \tag{4.25}$$

Από την Εξίσωση 4.24 και την Εξίσωση 4.25, ισχύει (για την προσέγγιση των Li-Ogihara-Zhou) ότι:

$$dis(F_\sigma, C_\sigma) \leq dis(F_\sigma, M_\sigma) \tag{4.26}$$

Οι εξισώσεις 4.20, 4.23 ανδ 4.26 δηλώνουν πως όσο πιο συμπαγής είναι η αναπαράσταση των στοιχειοσυνόλων στο πλέγμα (MFIs vs CFIs vs FIs), τόσο μεγαλύτερη είναι η υπολογιζόμενη απόσταση, και αυτό ισχύει για όλα τα εξεταζόμενα μέτρα ανομοιότητας.

4.5 Πειραματική αξιολόγηση

Για την αξιολόγηση των θεωρητικών αποτελεσμάτων που παρουσιάσαμε στις Ενότητες 4.4.1 και 4.4.2, πειραματιστήκαμε με τα διάφορα μέτρα ανομοιότητας που έχουν προταθεί (Parthasarathy-Ogihara, FOCUS, Li-Ogihara-Zhu) σε σύνολα δεδομένων από την αποθήκη συνόλων δεδομένων για το FIM πρόβλημα [23].

Για τα πειράματα χρησιμοποιήσαμε τόσο συνθετικά όσο και πραγματικά σύνολα δεδομένων που χαρακτηρίζονται επιπλέον και με βάση την πυκνότητά τους ως πυκνά ή αραιά. Τα χαρακτηριστικά των συνόλων δεδομένων απεικονίζονται στον Πίνακα 4.2. Για την εξαγωγή των συνόλων FI, CFI, MFI χρησιμοποιήσαμε το πρόγραμμα $MAFIA$ [14].

σύνολο δεδομένων	# συναλλαγών	# στοιχείων	| μέση συναλλαγή|	τύπος	τύπος
T10I4Δ100K	100,000	1,000	10	αραιό	συνθετικό
chess	3,196	76	37	πυκνό	πραγματικό
connect	67,557	130	43	πυκνό	πραγματικό

Πίνακας 4.2: Τα χαρακτηριστικά των συνόλων δεδομένων

Για την προσέγγιση του FOCUS (βλέπε Ενότητα 4.3.2), χρησιμοποιήσαμε το άνω όριο της ανομοιότητας όπως πρωτάθηκε από τους συγγραφείς, χωρίς να ξαναρωτήσουμε τα πραγματικά σύνολα δεδομένων. Η απόφαση αυτή έχει να κάνει με το γεγονός ότι, όπως έχουμε ήδη αναφέρει, ο στόχος μας είναι να δούμε κατά πόσο τα πρότυπα (τα συχνά στοιχειοσύνολα στην περίπτωσή μας) μπορούν να χρησιμοποιηθούν για τη σύγκριση των πρωτογενών δεδομένων. Για την προσέγγιση των Parthasarathy-Ogihara (βλέπε Ενότητα 4.3.1), χρησιμοποιήσαμε την τιμή $\theta = 1$, θεωρώντας συνεπώς ότι η δομική και η ποσοτική συνιστώσα συνεισφέρουν εξίσου στο τελικό σκορ ανομοιότητας.

4.5.1 Ανομοιότητα στο χώρο των πρωτογενών δεδομένων και στο χώρο των προτύπων

Η πρώτη ομάδα πειραμάτων αξιολογεί το επιχείρημα ότι η ανομοιότητα στο χώρο των προτύπων μπορεί να χρησιμοποιηθεί για την αποτίμηση της ομοιότητας στο χώρο των πρωτογενών δεδομένων. Πιο συγκεκριμένα, επιλέγουμε μία δημοφιλή

αναπαράσταση για τα στοχειοσύνολα του πλέγματος, τα FI, και ένα συγκεκριμένο $minSupport$ κατώφλι (σ) για την εξαγωγή τους, ενώ τροποποιούμε το σύνολο δεδομένων προσθέτοντας διαφορετικές τιμές θορύβου. Στη συνέχεια, συγκρίνουμε την ανομοιότητα που υπολογίζουμε στον χώρο των FI με την ανομοιότητα που υπάρχει (λόγω της προσθήκης του θορύβου) στο χώρο των πρωτογενών δεδομένων.

Το σενάριο έχει ως εξής: Ξεκινώντας με ένα αρχικό σύνολο δεδομένων D και με ένα συγκεκριμένο $minSupport$ κατώφλι σ, εξάγουμε το $F_\sigma(D)$. Στη συνέχεια, σε κάθε βήμα, τροποποιούμε έναν αύξοντα αριθμό συναλλαγών $(0\%, 5\%, \ldots, 50\%)$ του D. Η επιλογή των προς τροποποίηση συναλλαγών γίνεται με τυχαίο τρόπο, και για κάθε συναλλαγή που επιλέγεται τροποποιούμε ένα σταθερό ποσοστό (50%) των στοιχείων της. Η τροποποίησή μας συνίσταται στην αντικατάσταση των τιμών των στοιχείων με 0 (σε ένα στάδιο προεπεξεργασίας μετατρέψαμε τα σύνολα δεδομένων σε δυαδική μορφή). Συνεπώς, τα παραγώμενα σύνολα προτύπων $F_\sigma(D_{p\%})$ αποτελούν υποσύνολα των αρχικών συνόλων προτύπων $F_\sigma(D_{0\%})$. Στη συνέχεια, συγκρίναμε τα σύνολα προτύπων με θόρυβο $F_\sigma(D_{p\%})$ με το αρχικό (“καθαρό”) σύνολο προτύπων $F_\sigma(D_{0\%})$.

Όσον αφορά στη δημιουργία των συνόλων προτύπων, χρησιμοποιήσαμε $\sigma = 80\%$ για το σύνολο δεδομένων $D = chess$ και $\sigma = 0.5\%$ για το σύνολο δεδομένων $D = T10I4D100K$. Η επιλογή αυτών των παραμέτρων έγινε με βάση την αναλυση που παρουσιάστηκε στην [82]. Τα αποτελέσματα παρατίθενται στην Εικόνα 4.4.

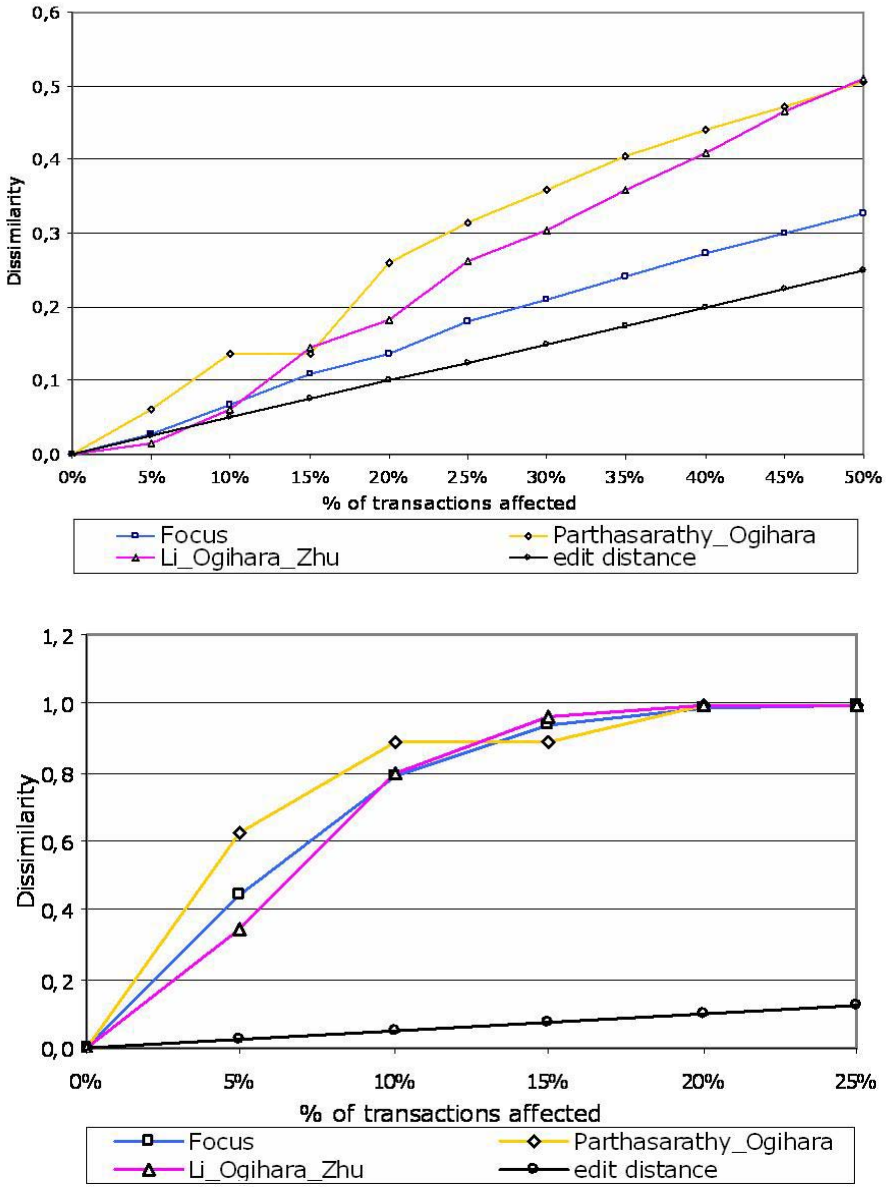
Όπως φαίνεται στις εικόνες αυτές, τόσο για το πυκνό όσο και για το αραιό σύνολο δεδομένων, καθώς περισσότερος θόρυβος προστίθεται στο αρχικό σύνολο δεδομένων, η απόσταση μεταξύ του αρχικού συνόλου προτύπων και του νέου (με θόρυβο) συνόλου προτύπων γίνεται μεγαλύτερη, και αυτό ισχύει για όλα τα προτεινόμενα μέτρα ανομοιότητας. Επιπλέον, αν θεωρήσουμε πως η απόσταση στο χώρο των πρωτογενών δεδομένων είναι η *απόσταση επεξεργασίας* (edit distance) μεταξύ του αρχικού και του “πειραγμένου” συνόλου δεδομένων, μπορούμε να δούμε πως κανένα από τα προτεινόμενα μέτρα δεν διατηρεί στο χώρο των προτύπων την απόσταση που εμφανίζεται στο χώρο των πρωτογενών δεδομένων.

Μία συγκριτική μελέτη των εικόνων αυτών δείχνει πως η επίδραση του θορύβου είναι πιο καταστροφική στην περίπτωση των πυκνών συνόλων δεδομένων όπως το chess. Στην Εικόνα 4.4 (κάτω), η ανομοιότητα αυξάνεται απότομα στη μέγιστη τιμής 1. Αυτό μπορεί να οφείλεται στο γεγονός ότι μικρές αλλαγές σε ένα πυκνό σύνολο δεδομένων μπορεί να οδηγήγουν σε πλέγματα στοιχειοσυνόλων που διαφέρουν σημαντικά μεταξύ τους. Από την άλλη, αυτό δεν ισχύει στην περίπτωση των αραιών συνόλων δεδομένων όπως το T10I4D100K το οποίο εμφανίζεται πιο εύρωστο στην αύξηση του θορύβου στο σύνολο δεδομένων, όπως φαίνεται στην Εικόνα 4.4 (επάνω).

4.5.2 Επίδραση του κατωφλίου $minSupport$

Στην ενότητα αυτή, αξιολογούμε πειραματικά την επίδραση του κατωφλίου $minSupport$ στην υπολογιζόμενη ανομοιότητα.

Το σενάριο έχει ως εξής: Για κάθε σύνολο δεδομένων D, σταθεροποιήσαμε ένα αρχικό $minSupport$ κατώφλι σ και εν συνεχεία μεταβάλλαμε την αύξηση δ στην τιμή του $minSupport$ στο εύρος τιμών $\delta_0, \delta_1, \ldots, \delta_n (\sigma + \delta_i \leq 1)$. Στη συνέχεια, εξάγαμε τα αντίστοιχα σύνολα από FI για τις διάφορες τιμές $minSupport$, $\sigma + \delta_0, \sigma + \delta_1, \ldots, \sigma + \delta_n$. Μετά συγκρίναμε τα σύνολα $FI_{\sigma+\delta_i}$ με το αρχικό σύνολο $FI_{\sigma+\delta_0}$. Η επιλογή των τιμών των παραμέτρων σ, δ για κάθε σύνολο δεδομένων



Σχήμα 4.4: Επίδραση του θορύβου του συνόλου δεδομένων στην ανομοιότητα των ΦΙ: $D = T10I4D100K$, $\sigma = 0.5\%$ (επάνω), $D = chess$, $\sigma = 80\%$ (κάτω).

έγινε με βάση την ανάλυση που παρουσιάστηκε στην [82]. Δεδομένου ότι η ανάλυσή μας δεν εξαρτάται από συγκεκριμένες τιμές του κατωφλίου $minSupport$, επιλέξαμε τιμές που οδηγούν σε λογικό αριθμό προτύπων. Έτσι, για το σύνολο δεδομένων $D = T10I4D100K$ δατασετ, επιλέξαμε αρχική τιμή $\sigma = 0,5\%$ και τιμές $0\%, 0.5\%, \ldots, 4.5\%$ για την αύξηση δ του $minSupport$ κατωφλίου. Για το σύνολο δεδομένων $D = chess$, επιλέξαμε αρχική τιμή $\sigma = 90\%$ και τιμές $0\%, 1\%, \ldots, 9\%$

για την αύξηση δ του $minSupport$ κατωφλίου.

Τα αποτελέσματα παρουσιάζονται στην Εικόνα 4.5. Όπως φαίνεται από τις εικόνες αυτές, όσο μεγαλύτερη είναι η αύξηση στην τιμή του $minSupport$ κατωφλίου δ, τόσο μεγαλύτερη είναι η ανομοιότητα που υπολογίζεται, και αυτό ισχύει για όλα τα προτεινόμενα μέτρα ανομοιότητας.

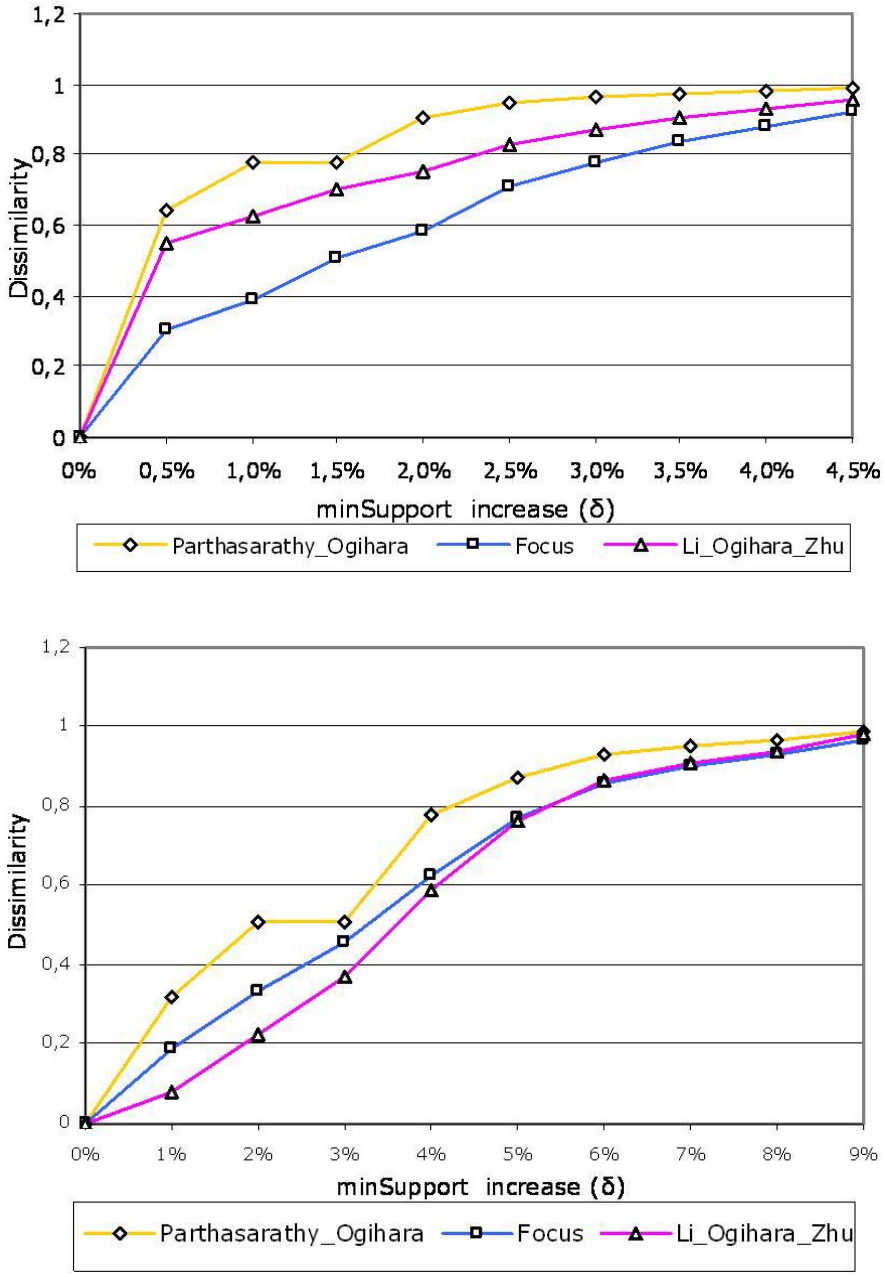
Λίγα παραπάνω σχόλια στα πειράματα: Τόσο η προσέγγιση των Parthasarathy-Ogihara όσο και η προσέγγιση του FOCUS βασίζονται σε κάποιο είδος 1-1 ταιριάσματος μεταξύ των (ίδιας δομής) στοιχειοσυνόλων των δύο συνόλων. Οι προσεγγίσεις αυτές διαφέρουν στο γεγονός ότι οι Parthasarathy-Ogihara λαμβάνουν υπόψη μόνο τα στοιχειοσύνολα που εμφανίζοντια στην τομή των δύο συνόλων ($A \cap B$), ενώ ο FOCUS λαμβάνει επίσης υπόψη και τα στοιχειοσύνολα που εμφανίζονται στα σύνολα $A - A \cap B$ και $B - A \cap B$, δηλαδή λαμβάνει υπόψη τα στοιχειοσύνολα που εμφανίζονται στην ένωση $A \cup B$. Αν ένα στοιχειοσύνολο εμφανίζεται σε ένα μόνο από τα δύο σύνολα, οι Parthasarathy-Ogihara θα αυξήσουν το συνολικό σκορ ανομοιότητας κατά 1 (σε απόλυτη τιμή), ενώ ο FOCUS θα αυξήσει το συνολικό σκορ τόσο όσο είναι η τιμή της υποστήριξης του συγκεκριμένου στοιχειοσυνόλου. Αυτός είναι ο λόγος που ο FOCUS υπολογίζει μικρότερες τιμές ανομοιότητας σε σχέση με τους Parthasarathy-Ogihara. Από την άλλη, οι Li-Ogihara-Zhou ακολουθούν μία διαφορετική λογική σε σχέση με τους Parthasarathy-Ogihara και FOCUS. Συγκεκριμένα, υιοθετούν ένα είδος $M - N$ ταιριάσματος μεταξύ των στοιχειοσυνόλων των δύο συνόλων και επίσης λαμβάνουν υπόψη και περιπτώσεις μερικής ομοιότητας μεταξύ των στοιχειοσυνόλων.

Παρατηρούμε επίσης μία διαφορά στη συμπεριφορά των διαφόρων μέτρων ανάλογα με το αν το σύνολο δεδομένων είναι πυκνό ή αραιό. Η επίδραση της παραμέτρου $minSupport$ στην περίπτωση των πυκνών συνόλων δεδομένων (όπως το *chess*) είναι πιο ομαλή σε σχέση με τα αραιά σύνολα δεδομένων (όπως το *T10I4D100K*). Αυτό οφείλεται πιθανόν στο γεγονός ότι σε ένα πυκνό σύνολο δεδομένων τα στοιχειοσύνολα διαχέονται πιο ομαλά στα διάφορα επίπεδα του $minSupport$ σε σχέση με κάποιο αραιό σύνολο δεδομένων. Αυτό επιβεβαιώνεται στην Εικόνα 4.6 όπου παρουσιάζεται η κατανομή του πλήθους των στοιχειοσυνόλων στα διάφορα επίπεδα του $minSupport$, για κάθε σύνολο δεδομένων.

Συνοψίζοντας, τα πειράματα στην υποενότητα αυτή επαληθεύουν την θεωρητική μας ανάλυση όσον αφορά στην εξάρτηση της ανομοιότητας δύο συνόλων στοιχειοσυνόλων από το $minSupport$ κατώφλι που χρησιμοποιήθηκε για την εξαγωγή τους. Πράγματι, καθώς το $minSupport$ αυξάνεται, η ανομοιότητα αυξάνεται. Γενικεύοντας αυτό το αποτέλεσμα, μπορούμε να πούμε πως όσο πιο "επιλεκτικό" είναι το $minSupport$ κατώφλι, τόσο λιγότερη πληροφορία περιέχει το σύνολο στοιχειοσυνόλων σε σχέση με το σύνολο πρωτογενών δεδομένων από το οποίο έχει εξαχθεί. Συνεπώς, όταν συγκρίνουμε δύο σύνολα δεδομένων με βάση τα αντίστοιχα σύνολα προτύπων, η υπολογιζόμενη ανομοιότητα εξαρτάται από το $minSupport$ κατώφλι που χρησιμοποιήθηκε για την εξαγωγή των συνόλων από στοιχειοσύνολα.

4.5.3 Επίδραση του επιπέδου συμπίεσης του πλέγματος

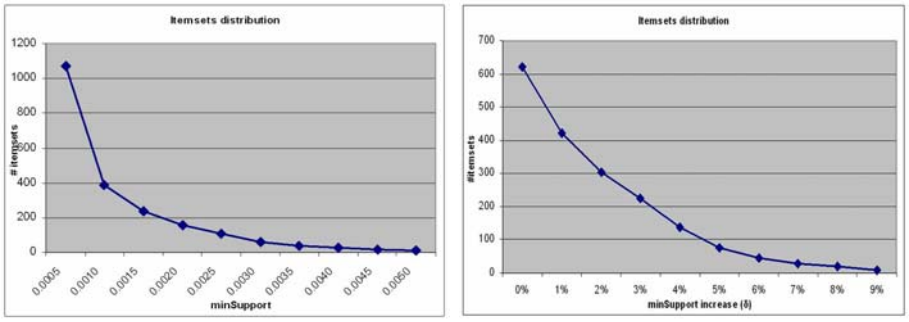
Στην ενότητα αυτή, μελετάμε την επίδραση των διαφορετικών αναπαραστάσεων των στοιχειοσύνολων του πλέγματος, στο αποτέλεσμα της σύγκρισης. Πιο συγκεκριμένα, αρχικοποιούμε το $minSupport$ κατώφλι σ σε μία συγκεκριμένη τιμή και υπολογίζουμε την ανομοιότητα μεταξύ των διαφόρων αναπαραστάσεων του πλέγματος για διάφορα επίπεδα θορύβου, δηλαδή, $dis(F_\sigma(D_{p\%}), C_\sigma(D_{p\%}))$, $dis(F_\sigma(D_{p\%}), M_\sigma(D_{p\%}))$.



Σχήμα 4.5: Επίδραση της αύξησης δ του $minSupport$ στην ανομοιότητα των συνόλων FI: $D = T10I4D100K$, $\sigma = 0.5\%$ (τοπ), $D = chess$, $\sigma = 90\%$ (βοττομ).

Τα αποτελέσματα για τις συγκρίσεις $FI-CFI$, $FI-MFI$ παρατίθενται στο Σχήμα 4.7.

Οι γραφικές αυτές δείχνουν την εξάρτηση των αποτελεσμάτων της σύγκρισης



(α) (β)

Σχήμα 4.6: Η κατανομή των στοιχειοσυνόλων για διάφορες τιμές υποστήριξης (α) $D = T10I4D100K$, $\sigma = 0.5\%$ και (β) $D = chess$, $\sigma = 90\%$

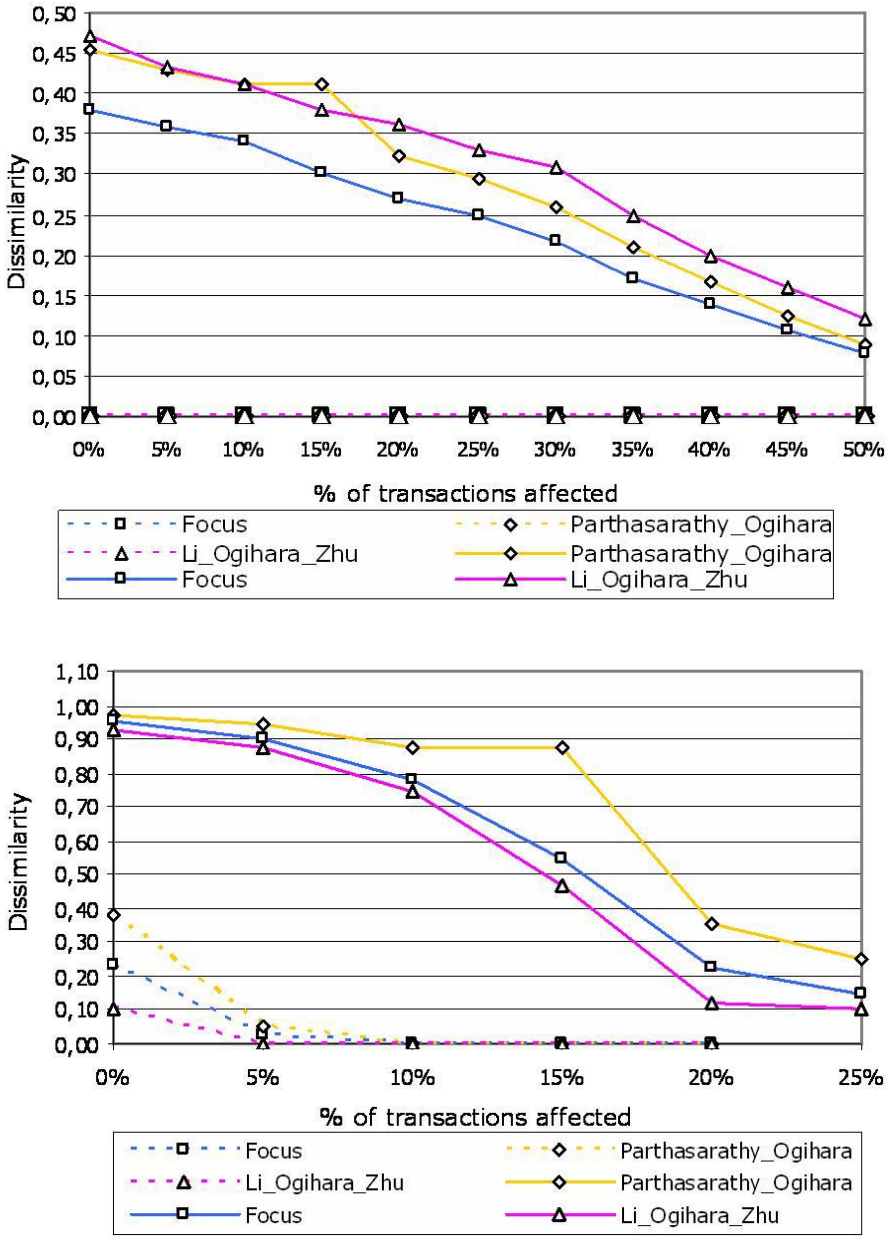
από την αναπαράσταση των στοιχειοσυνόλων του πλέγματος. Όπως φαίνεται από τις εικόνες αυτές, τα σύνολα CFI μπορούν να πιάσουν πολύ καλά τη συμπεριφορά των συνόλων FI ενώ τα σύνολα MFI όχι, και αυτό ισχύει και για τα δύο σύνολα δεδομένων.

Ωστόσο η διαφορά ανάμεσα στα αποτελέσματα της σύγκρισης των $FI - CFI$ και $FI - MFI$ φαίνεται να εξαρτάται από τα χαρακτηριστικά του συνόλου δεδομένων (αραιά ή πυκνά). Πιο συγκεκριμένα, στην περίπτωση των αραιών συνόλων δεδομένων όπως το $T10I4D100K$, τα CFI καταφέρνουν να πιάσουν πλήρως την συμπεριφορά των FI σε κάθε επίπεδο θορύβου ενώ τα MFI ομαλά προσεγγίζουν τα FI καθώς το σύνολο δεδομένων γίνεται πιο θορυβώδες (βλέπε Σχήμα 4.7, επάνω). Στην περίπτωση των πυκνών συνόλων δεδομένων όπως το $\varsigma\eta\varepsilon\sigma\varsigma$, τα CFI εξακολουθούν να παραμένουν κοντά στα FI παρόλο που δεν πιάνουν πλήρως τη συμπεριφορά τους (βλέπε Σχήμα 4.7, κάτω). Όπως φαίνεται στις εικόνες αυτές, οι διαφορές $FI - CFI$, $FI - MFI$ είναι πιο ισχυρές στην περίπτωση των πυκνών συνόλων δεδομένων.

Συνοψίζοντας, τα πειράματα στην υποενότητα αυτή δείχνουν πως η αναπαράσταση των στοιχειοσυνόλων του πλέγματος επηρεάζει το αποτέλεσμα της σύγκρισης, επιβεβαιώνοντας τη θεωρητική μας ανάλυση. Επιπλέον, φαίνεται πως τα CFI προσεγγίζουν πολύ καλά τη συμπεριφορά των FI σε αντίθεση με τα MFI. Φαίνεται πως όσο πιο συμπαγής είναι η αναπαράσταση του συνόλου των στοιχειοσυνόλων (προτύπων, γενικότερα), τόσο λιγότερη πληροφορία περιέχει αυτός ο χώρος σε σχέση με τον αρχικό χώρο των πρωτογενών δεδομένων από τα οποία τα πρότυπα έχουν εξαχθεί.

4.6 Συμπεράσματα

Σε αυτό το κεφάλαιο, μελετήσαμε κατά πόσο η σύγκριση μεταξύ συνόλων από συχνά στοιχειοσύνολα θα μπορούσε να αποτελέσει μέτρο σύγκρισης για τα σύνολα πρωτογενών δεδομένων από τα οποία εξήχθησαν τα στοιχειοσύνολα αυτά. Παρουσιάσαμε τις παραμέτρους που επηρεάζουν το πρόβλημα και συγκεκριμένα, το ελάχιστο κατώφλι υποστήριξης $minSupport$ που χρησιμοποιείται για την εξαγωγή των στοιχειοσυνόλων και το επίπεδο συμπίεσης της αναπαράστασης του πλέγματος (συχνά στοιχειοσύνολα - FI, κλειστά συχνά στοιχειοσύνολα - CFI και μέγιστα



Σχήμα 4.7: Επίδραση του θορύβου στην ανομοιότητα των $FI-CFI$ (διακεκομένες γραμμές), $FI-MFI$ (ενιαίες γραμμές): $D = T10I4D100K$, $\sigma = 0.5\%$ (τοπ), $D = chess$, $\sigma = 80\%$ (βοττομ).

συχνά στοιχειοσύνολα - MFI). Τόσο η θεωρητική όσο και η πειραματική ανάλυση έδειξαν πως όσο πιο περιοριστικές είναι οι παράμετροι της εξόρυξης τόσο πιο μεγάλη είναι η υπολογιζόμενη απόσταση, και αυτό ισχύει για όλες τα προτεινόμενα μέτρα ανομοιότητας.

Το αποτέλεσμα αυτό δηλώνει πως το να χρησιμοποιήσουμε τη σύγκριση μεταξύ στοιχειοσυνόλων για τη σύγκριση των πρωτογενών συνόλων δεδομένων δεν είναι τόσο προφανές, αλλά είναι υποκειμενικό όσον αφορά στις παραμέτρους που χρησιμοποιούνται για την εξόρυξη των προτύπων.

Μία πιθανή εξήγηση για την παρόμοια συμπεριφορά των διαφόρων μέτρων ανομοιότητας που έχουν προταθεί όσον αφορά στις παραμέτρους της εξόρυξης γνώσης είναι ότι όλα τα μέτρα ακολουθούν ένα γενικό σχήμα (βλέπε Ενότητα 4.3.4). Πιο συγκεκριμένα, όλα τα μέτρα βασίζονται είτε στην ένωση είτε στην τομή των προς σύγκριση συνόλων από στοιχειοσύνολα. Ωστόσο, η επιλογή της τιμής του κατωφλίου $minSupport$ και του επιπέδου συμπίεσης του πλέγματος (FI, CFI ή MFI) καθορίζει το σύνολο από στοιχειοσύνολα που θα προκύψει. Επίσης, όπως φαίνεται και από τα πειράματα, μία μικρή αλλαγή στο σύνολο δεδομένων μπορεί να οδηγήσει σε πολύ διαφορετικά σύνολα στοιχειοσυνόλων. Όλα αυτά τα στοιχεία, δείχνουν πως όταν κάποιος χρησιμοποιεί τα σύνολα των προτύπων (συχνά στοιχειοσύνολα στην περίπτωσή μας) για τη σύγκριση των συνόλων δεδομένων θα πρέπει να λαμβάνει υπόψη τις παραμέτρους της εξόρυξης. Μία λογική επιλογή θα ήταν να χρησιμοποιήσουμε τα σύνολα στοιχειοσυνόλων που εξάγονται στο χαμηλότερο κατώφλι υποστήριξης $minSupport$ και χρησιμοποιούν το χαμηλότερο επίπεδο συμπίεσης όσον αφορά στην αναπαράσταση του πλέγματος (δηλαδή, τα σύνολα FI). Αυτά τα σύνολα είναι πιο πιθανό να μεταφέρουν το μεγαλύτερο ποσοστό πληροφορίας που υπάρχει στο σύνολο των πρωτογενών δεδομένων. Ειδικά για το επίπεδο συμπίεσης του πλέγματος, όπως φάνηκε από τα παραδείγματα, τα σύνολα CFI είναι πολύ κοντά στα σύνολα FI, συνεπώς θα μπορούσε κανείς να χρησιμιποιήσει τα σύνολα CFI για τη σύγκριση. Ωστόσο, αυτή η επιλογή εξαρτάται επίσης και από τα χαρακτηριστικά του συνόλου των δεδομένων (πυκνό ή αραιό). Όπως φάνηκε στα πειράματα, τα πυκνά σύνολα δεδομένων είναι λιγότερο εύρωστα στις παραμέτρους της εξόρυξης σε σχέση με τα αραιά σύνολα δεδομένων.

Πρώιμες εκδόσεις αυτής της δουλειάς εμφανίζονται στις [54, 56], ενώ μία εκτεταμένη έκδοση έχει υποβληθεί στην [62]

4.7 Ανοιχτά θέματα

Ένα ανοιχτό θέμα είναι η εύρεση ενός μέτρου ανομοιότητας το οποίο θα είναι εύρωστο στις παραμέτρους της Εξόρυξης Γνώσης και θα διατηρεί καλύτερα τα χαρακτηριστικά του χώρου των πρωτογενών δεδομένων (data space) στο χώρο των προτύπων (pattern space). Από την παρούσα μελέτη φαίνεται πως μία $M-N$ σύγκριση μεταξύ των στοιχειοσυνόλων των δύο συνόλων είναι πιο σταθερή σε σχέση με μία $1-1$ σύγκριση. Επίσης, τα αποτελέσματα της σύγκρισης είναι πιο σταθερά όταν κάποιος λαμβάνει υπόψη τα στοιχειοσύνολα που εμφανίζονται στην ένωση των δύο συνόλων αντί της τομής. Ωστόσο, εκτός από το να εξετάζουμε τα στοιχειοσύνολα που εμφανίζονται στην ένωση ή την τομή των δύο συνόλων, υπάρχουν και άλλες εναλλακτικές προσεγγίσεις. Για παράδειγμα, θα μπορούσε κανείς να ακολουθήσει μία προσέγγιση φάσματος (spectrum like approach) εξετάζοντας τα στοιχειοσύνολα με συγκεκριμένο πλήθος στοιχείων (k), π.χ., όλα τα $2-$στοιχειοσύνολα, αντί όλων των στοιχειοσυνόλων. Μία τέτοια προσέγγιση θα μπορούσε να αποτελεί μία πιο λογική και δίκαιη αντανάκλαση του αρχικού συνόλου δεδομένων.

Κεφάλαιο 5

Σύγκριση Δέντρων Απόφασης (και Συνόλων Δεδομένων Κατηγοριοποίησης) - Σημασιολογική Ομοιότητα Δέντρων Απόφασης

Στο κεφάλαιο αυτό, παρουσιάζουμε ένα γενικό πλαίσιο για την αποτίμηση της ομοιότητας μεταξύ δέντρων απόφασης (decision trees) και συνόλων δεδομένων κατηγοριοποίησης (classification datasets). Η προσέγγισή μας βασίζεται στην τμηματοποίηση που δημιουργεί ένα δέντρο απόφασης πάνω στο χώρο των γνωρισμάτων του προβλήματος. Επιδεικνύουμε την χρησιμότητα και την εφαρμοσιμότητα του πλαισίου μας στην αποτίμηση της *σημασιολογικής ομοιότητας* (semantic similarity) δέντρων απόφασης που έχουν εξαχθεί από διαφορετικά υποσύνολα συνόλων δεδομένων κατηγοριοποίησης. Αξιολογούμε την απόδοση του προτεινόμενου μέτρου σημασιολογικής ομοιότητας σε σχέση με το μέτρο της *εμπειρικής σημασιολογικής ομοιότητας*, το οποίο υπολογίζεται με βάση ανεξάρτητα hold out σύνολα ελέγχου.

Το κεφάλαιο έχει οργανωθεί ως εξής: Η Ενότητα 5.1 αποτελεί μία εισαγωγή στο πρόβλημα. Στην Ενότητα 5.2, περιγράφουμε κάποιες βασικές έννοιες σε δέντρα απόφασης, οι οποίες είναι απαραίτητες για την περαιτέρω κατανόηση αυτού του κεφαλαίου. Στην Ενότητα 5.3, περιγράφουμε το πλαίσιο ομοιότητας που προτείνουμε. Στην Ενότητα 5.4, αξιολογούμε πειραματικά το προτεινόμενο μέτρο σημασιολογικής ομοιότητας δέντρων απόφασης. Στην Ενότητα 5.5, παρουσιάζουμε τις σχετικές εργασίες. Στην Ενότητα 5.6 συνοψίζουμε τα βασικά σημεία του κεφαλαίου, ενώ στην Ενότητα 5.7 συζητάμε σχετικά ανοιχτά ερευνητικά θέματα.

Λέξεις κλειδιά δέντρα απόφασης, μέτρα ομοιότητας, σημασιολογική ομοιότητα, σύνολα δεδομένων κατηγοριοποίησης.

5.1 Εισαγωγή

Τα δέντρα απόφασης (decision trees - DT) αποτελούν ένα από τα πιο δημοφιλή παραδείγματα μάθησης στον τομέα της Εξόρυξης Γνώσης λόγω ενός πλήθους ελκυστικών ιδιοτήτων, όπως η κλιμάκωση σε μεγάλα σύνολα δεδομένων και η σχετική ευκολία στην ερμηνεία τους από τους τελικούς χρήστες, με την προϋπόθεση βέβαια πως το μέγεθός τους δεν υπερβαίνει συγκεκριμένα όρια.

Από την άλλη, τα δέντρα απόφασης είναι επίσης πολύ γνωστά για την *αστάθειά* (instability) τους. Μικρές αλλαγές στο σύνολο δεδομένων εκπαίδευσης (training set) μπορεί να οδηγήσουν σε εντελώς διαφορετικά δέντρα απόφασης, τα οποία μπορεί να περιέχουν διαφορετικές συνθήκες ελέγχου στα γνωρίσματα πρόβλεψης αλλά και διαφορετικά γνωρίσματα πρόβλεψης (predictive attributes). Να σημειώσουμε επίσης πως δύο δέντρα απόφασης, παρόλο που μπορεί να διαφέρουν ως προς τη δομή τους, μπορεί να περιγράφουν την ίδια έννοια, δηλαδή, μπορεί να είναι *σημασιολογικά παρόμοια* ή ακόμα και ταυτόσημα. Διαισθητικά, δύο δέντρα απόφασης θα πρέπει να έχουν σημασιολογική ομοιότητα αν έχουν προκύψει από σύνολα δεδομένων που προέρχονται από την ίδια γεννήτρια κατανομή.

Δύο δέντρα μπορεί να εμφανίζουν σημασιολογική ομοιότητα ακόμα και αν η δομή τους είναι διαφορετική. Μία τέτοια συμπεριφορά μπορεί να οφείλεται στο γεγονός ότι υπάρχουν επικαλυπτόμενα ή ισοδύναμα γνωρίσματα πρόβλεψης και γενικότερα, μία έννοια μπορεί να περιγραφεί με διαφορετικούς τρόπους. Συνεπώς, για να μπορούμε να υπολογίζουμε το βαθμό σημασιολογικής ομοιότητας δύο δέντρων απόφασης, χρειαζόμαστε ένα μέτρο της σημασιολογικής ομοιότητας των εννοιών που περιγράφουν.

Υπάρχει ένα πλήθος εφαρμογών όπου είναι απαραίτητη η ύπαρξη ενός μέτρου σημασιολογικής ομοιότητας μεταξύ των δέντρων απόφασης. Η πιο σημαντική ίσως εφαρμογή είναι η δυνατότητα να αξιολογήσουμε κατά πόσο οι διαφορές που παρατηρούνται μεταξύ δύο δέντρων απόφασης, τα οποία προέρχονται από διαφορετικά σύνολα δεδομένων εκπαίδευσης (που όμως θα πρέπει να προέρχονται από την ίδια γεννήτρια κατανομή), είναι μόνο δομικές και δεν ανταποκρίνονται σε σημασιολογικές διαφορές ή αν οι έννοιες που αναπαριστούν τα δέντρα είναι όντως διαφορετικές. Στη δεύτερη περίπτωση, είναι σημαντικό να μπορούμε επιπλέον να ποσοτικοποιούμε αυτή τη διαφορά.

Επιπλέον, η ύπαρξη ενός μέτρου σημασιολογικής ομοιότητας μεταξύ μοντέλων ταξινόμησης επιτρέπει την εφαρμογή κλασικών εργασιών Εξόρυξης Γνώσης πάνω στα ίδια τα μοντέλα (meta-mining) αντί των πρωτογενών δεδομένων (όπως συμβαίνει παραδοσιακά). Για παράδειγμα, η σημασιολογική ομοιότητα θα μπορούσε να χρησιμοποιηθεί για την συσταδοποίηση των δέντρων απόφασης μιας τράπεζας (τα οποία αναφέρονται π.χ. στο πρόβλημα της χορήγησης δανείων), όπου κάθε δέντρο αντιστοιχεί σε ένα υποκατάστημα της τράπεζας. Με τον τρόπο αυτό, η τράπεζα μπορεί να χαράξει στρατηγικές για κάθε συστάδα υποκαταστημάτων και όχι για κάθε κατάστημα ξεχωριστά.

Επίσης, στην περίπτωση δυναμικών δεδομένων όπως τα ρεύματα δεδομένων (data streams), η ομοιότητα θα μπορούσε να χρησιμοποιηθεί προκειμένου να παρακολουθεί κανείς την εξέλιξη των δέντρων απόφασης στο χρόνο. Ένα κρίσιμο ερώτημα στην περίπτωση αυτή είναι αν η έννοια που αναπαριστά το δέντρο απόφασης παραμένει (περίπου) ίδια ή αν υπάρχουν μετατοπίσεις στον πληθυσμό (concept shift). Έχοντας στη διάθεσή της τέτοιου είδους πληροφορίες μία τηλεπικοινωνιακή εταιρία θα μπορούσε, για παράδειγμα, να προσαρμόσει εγκαίρως την πολιτική χρέωσης που ακολουθεί.

Η ομοιότητα θα μπορούσε να χρησιμοποιηθεί επίσης προκειμένου να μελετήσουμε την επίδραση των διαφόρων παραμέτρων μάθησης, όπως το επίπεδο κλαδέματος ή ο αλγόριθμος μάθησης, στα δέντρα απόφασης. Μία άλλη εφαρμογή είναι η σύγκριση ενός δέντρου απόφασης που εξάγεται από τα δεδομένα με ένα πρότυπο δέντρο απόφασης (βλέπε Εικόνα 1.3), έτσι ώστε να ξέρουμε πόσο απέχει το εξαχθέν (πραγματικό) δέντρο απόφασης από το αναμενόμενο (πιθανόν τεχνητό) μοντέλο.

Στο κεφάλαιο αυτό, προτείνουμε ένα γενικό πλαίσιο αποτίμηση της ομοιότητας, το οποίο περιλαμβάνει ως ειδικές περιπτώσεις: α) τον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης και β) τον υπολογισμό της ομοιότητας μεταξύ συνόλων δεδομένων κατηγοριοποίησης με βάση τις διαφορετικές κατανομές που διέπουν τα σύνολα αυτά, και συγκεκριμένα με βάση την *συνάρτηση κατανομής των γνωρισμάτων πρόβλεψης*, την *από κοινού συνάρτηση κατανομής των γνωρισμάτων πρόβλεψης και του γνωρίσματος κλάσης* και την *υπό-συνθήκη συνάρτηση κατανομής του γνωρίσματος κλάσης με βάση τα γνωρίσματα πρόβλεψης*. Το προτεινόμενο πλαίσιο βασίζεται στη σύγκριση των τμηματοποιήσεων που ορίζουν τα δέντρα απόφασης πάνω σε ένα δοθέν χώρο γνωρισμάτων, λαμβάνοντας υπόψη την κατανομή του χώρου δεδομένων πάνω σε αυτή την τμηματοποίηση. Ανάλογα με τη γνώση που έχουμε σχετικά με την συνάρτηση κατανομής που δημιούργησε τα σύνολα δεδομένων, προκύπτουν διαφορετικά στιγμιότυπα του μέτρου σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης. Για να αξιολογήσουμε το προτεινόμενο μέτρο *σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης*, το συγκρίνουμε με το μέτρο της εμπειρικής σημασιολογικής ομοιότητας (empirical semantic similarity), η οποία υπολογίζεται εφαρμόζοντας τα δέντρα απόφασης σε ανεξάρτητα σύνολα δεδομένων ελέγχου.

5.2 Βασικές έννοιες σε δέντρα απόφασης

Στην ενότητα αυτή, παρουσιάζουμε κάποιες βασικές έννοιες σε δέντρα απόφασης, οι οποίες είναι απαραίτητες για την κατανόηση αυτού του κεφαλαίου. Πιο γενικές πληροφορίες σχετικά με τα δέντρα απόφασης υπάρχουν στην Ενότητα 2.3.

Έστω ένα πρόβλημα κατηγοριοποίησης (classification problem) το οποίο περιγράφεται μέσω ενός διανύσματος από *γνωρίσματα πρόβλεψης* (predictive attributes) $A = (a_1, a_2, ..., a_m)$ και ένα *γνώρισμα κλάσης* (class attribute) C. Κάθε γνώρισμα a_i έχει πεδίο ορισμού $d(a_i)$, ενώ το πεδίο ορισμού για το γνώρισμα κλάσης είναι $d(C) = \{c_1, c_2, ..., c_k\}$, όπου k είναι το πλήθος των κλάσεων. Τα γνωρίσματα πρόβλεψης μπορεί είναι αριθμητικά (numerical), κατηγορηματικά (categorical) ή διατεταγμένα (ordinal). Συνήθως τα γνωρίσματα είναι αριθμητικά [50].

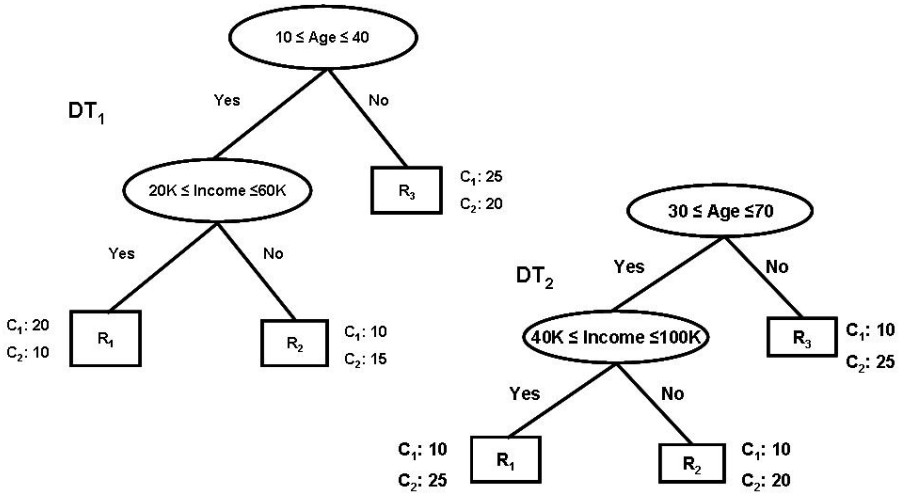
Το καρτεσιανό γινόμενο $S_A = d(a_1) \times d(a_2) \cdots \times d(a_m)$ περιγράφει το *χώρο γνωρισμάτων* (attribute space), ενώ το καρτεσιανό γινόμενο $S_{(A,C)} = S_A \times d(C)$ ορίζει το *χώρο γνωρισμάτων-κλάσης* (attribute-class space).

Ένα δέντρο απόφασης αποσκοπεί στο να μάθει μία *συνάρτηση πρόβλεψης* (prediction function) $f : S(A) \rightarrow dom(C)$. Για το λόγο αυτό, ο αλγόριθμος μάθησης παίρνει ως είσοδο ένα σύνολο δεδομένων D από στιγμιότυπα του προβλήματος, γνωστό ως *σύνολο δεδομένων εκπαίδευσης* (training set). Το σύνολο εκπαίδευσης D θα πρέπει να προέρχεται από την απο κοινού κατανομή $P(A, C)$ των γνωρισμάτων πρόβλεψης A και του γνωρίσματος κλάσης C, έτσι ώστε να είναι αντιπροσωπευτικό του συγκεκριμένου προβλήματος κατηγοριοποίησης. Σίγουρα, η επιλογή του D είναι κρίσιμη καθώς επηρεάζει την ικανότητα γενίκευσης του δέντρο απόφασης

πάνω σε μελλοντικά, προηγουμένως άγνωστα στιγμιότυπα του προβλήματος.

Έστω $U(A)$ είναι η *ομοιόμορφη κατανομή* (uniform distribution) πάνω στο χώρο γνωρισμάτων και $P(A) = \sum_C P(A, C)$ είναι η *οριακή κατανομή των γνωρισμάτων* (marginal distribution of attributes) που ορίζεται επίσης επίσης πάνω στο χώρο γνωρισμάτων.

Για καλύτερη κατανόηση, ας θεωρήσουμε ως παράδειγμα σύγκρισης τα δέντρα απόφασης της Εικόνας 5.1. Και τα δύο δέντρα αναφέρονται στο ίδιο πρόβλημα κατηγοριοποίησης, το οποίο περιγράφεται μέσω δύο γνωρισμάτων πρόβλεψης και ενός γνωρίσματος κλάσης. Τα γνωρίσματα πρόβλεψης είναι Age και $Income$ με πεδία ορισμού $(10 \leq Age \leq 70)$ και $(20K \leq Income \leq 100K)$, αντίστοιχα. Οι τιμές για το γνώρισμα κλάση C είναι: $C_1 = HighRisk, C_2 = LowRisk$.



Σχήμα 5.1: Δύο δέντρα απόφασης DT_1, DT_2

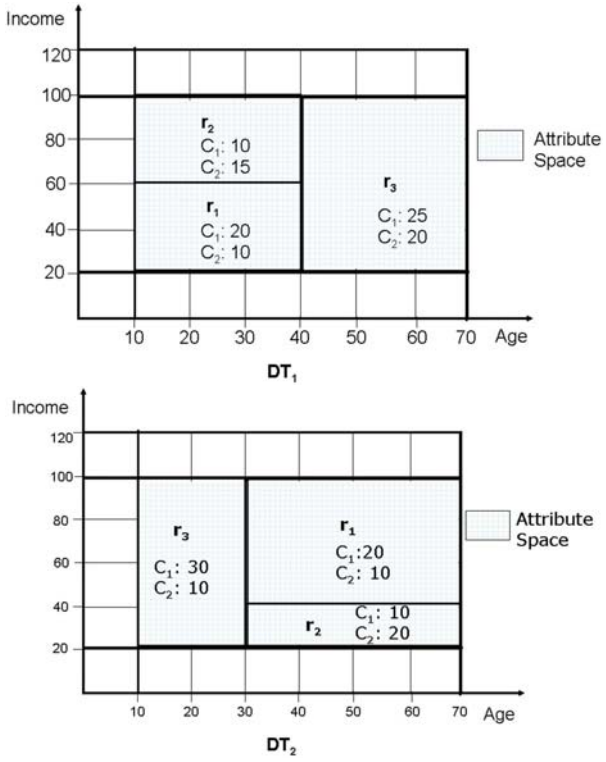
Ο Πίνακας 5.1 συνοψίζει τα σύμβολα που χρησιμοποιούνται σε αυτό το κεφάλαιο.

Σύμβολο	**Περιγραφή**
DTp	δέντρο απόφασης
R_{DT}	η τμηματοποίηση του DT
A	γνωρίσματα πρόβλεψης
C	γνώρισμα κλάσης
S_A	χώρος γνωρισμάτων
$S_{(A,C)}$	χώρος γνωρισμάτων - κλάσης
$U(A)$	ομοιόμορφη κατανομή στον S_A
$P(A)$	κατανομή των γνωρισμάτων στον S_A
$P(A, C)$	από κοινού κατανομή γνωρισμάτων και κλάσης
$P(C\|A)$	υπό συνθήκη κατανομή κλάσης βάσει των γνωρισμάτων

Πίνακας 5.1: Λίστα συμβόλων για το Κεφάλαιο 5

5.3 Αποτίμηση της ομοιότητας μέσω δέντρων απόφασης

Ένα δέντρο απόφασης DT, το οποίο προέρχεται από ένα σύνολο δεδομένων D, τμηματοποιεί το χώρο γνωρισμάτων σε ένα σύνολο μη επικαλυπτόμενων περιοχών $R_{DT} = \{r_i | i = 1 \dots |R_{DT}|\}$, μέσω των κόμβων-φύλλων του. Οι τμηματοποιήσεις των δέντρων απόφασης του παραδείγματός μας (βλέπε Εικόνα 5.1) παρατίθενται στην Εικόνα 5.2.



Σχήμα 5.2: Η τμηματοποίηση του DT_1 (επάνω), DT_2 (κάτω)

Η τμηματοποίηση R_{DT} που ορίζεται από το δέντρο απόφασης DT μπορεί να θεωρηθεί ως προσέγγιση της απο κοινού κατανομής των γνωρισμάτων πρόβλεψης και του γνωρίσματος κλάσης $P(A, C)$, με τη μορφή ενός ιστογράμματος (Ενότητα 5.3.1). Κάθε κουτί του ιστογράμματος αντιστοιχεί σε μία περιοχή της τμηματοποίησης, και συνεπώς, σε ένα κόμβο-φύλλο του δέντρου απόφασης. Διαφορετικά δέντρα απόφασης ορίζουν διαφορετικές τμηματοποιήσεις. Στην Ενότητα 5.3.2, περιγράφουμε πως μπορούμε να βρούμε την επικάλυψη των τμηματοποιήσεων δύο δέντρων απόφασης και πως μπορούμε να υπολογίσουμε τα στατιστικά της, για κάθε σύνολο δεδομένων, ανάλογα με το αν έχουμε πρόσβαση στα πρωτογενή σύνολα δεδομένων. Με βάση την επικάλυψη των τμηματοποιήσεων, ορίζουμε διάφορα μέτρα ομοιότητας μεταξύ δέντρων απόφασης και συνόλων δεδομένων κατηγοριοποίησης (Ενότητα 5.3.3).

5.3.1 Τμηματοποίηση δέντρου απόφασης

Κάθε περιοχή $r \in R_{DT}$ χαρακτηρίζεται από μία δομική συνιστώσα και από μία ποσοτική συνιστώσα που εξάγονται κατευθείαν από το δέντρο απόφασης.

5.3.1.1 Δομική συνιστώσα

Η *δομική συνιστώσα* (structure component) της περιοχής ορίζεται ως η συνένωση των συνθηκών ελέγχου των γνωρισμάτων κατά μήκος του αντίστοιχου μονοπατιού του δέντρου, το οποίο ξεκινάει από τη ρίζα του δέντρου και καταλήγει στο φύλλο που σχετίζεται με τη συγκεκριμένη περιοχή:

$$r.s = \{\wedge t_i(a_i), i = 1, ..., m\}$$

Οι συνθήκες ελέγχου είναι συνήθως αριθμητικές και μπορούν να εκφραστούν με τη μορφή: $t(a) = \min_a(r) \leq a \leq \max_a(r)$ συμβολίζοντας την ελάχιστη και τη μέγιστη τιμή του γνωρίσματος a στην περιοχή r.

Ας ορίσουμε επίσης το μήκος μίας συνθήκης ελέγχου σε ένα γνώρισμα a ως: $|t(a)| = \max_a(r) - \min_a(r)$ και το μήκος του πεδίου ορισμού του γνωρίσματος a ως: $|dom(a)| = \max_a - \min_a$. Να τονίσουμε εδώ πως αν ένα γνώρισμα a δεν περιλαμβάνεται στη δομική συνιστώσα $r.s$ ενός φύλλου, δηλαδή, δεν υπάρχει συνθήκη ελέγχου πάνω σε αυτό το γνώρισμα στο αντίστοιχο μονοπάτι του δέντρου, τότε η συνθήκη ελέγχου στο γνώρισμα αυτό είναι: $t(a) = \min_a \leq a \leq \max_a$, δηλαδή, το γνώρισμα a μπορεί να πάρει στην περιοχή r οποιαδήποτε τιμή από το πεδίο ορισμού του. Με τον τρόπο αυτό, η δομική συνιστώσα μιας περιοχής περιέχει συνθήκες ελέγχου πάνω σε όλα τα γνωρίσματα πρόβλεψης του προβλήματος κατηγοριοποίησης.

5.3.1.2 Ποσοτική συνιστώσα

Η *ποσοτική συνιστώσα* (measure component) μιας περιοχής ορίζεται ως το πλήθος των στιγμιοτύπων του συνόλου εκπαίδευσης που καταλήγουν στην περιοχή για κάθε κλάση του προβλήματος, και εξαρτάται από το σύνολο εκπαίδευσης D που χρησιμοποιήθηκε για την εκπαίδευση του δέντρου απόφασης:

$$r.m_D = [n_{c_1}, n_{c_2}, \ldots, n_{c_k}]$$

όπου $n_{c_i}, i = 1 \ldots k$ είναι το πλήθος των στιγμιοτύπων που καταλήγουν στην περιοχή r και ανήκουν στην κλάση c_i.

Ο συνολικός αριθμός στιγμιοτύπων στην περιοχή r, μπορεί εύκολα να υπολογιστεί αθροίζοντας το πλήθος των στιγμιούπων της r για κάθε κλάση του προβλήματος, συνεπώς το μέγεθος της δομικής συνιστώσας είναι:

$$|r.m_D| = \sum_{1 \leq i \leq k} n_{c_i}$$

Η κλάση που ανατίθεται στην περιοχή r, είναι η κλάση στην οποία ανήκει η πλειοψηφία των στιγμιοτύπων:

$$r.cl = \arg\max_{c_i} r.m_D$$

Για παράδειγμα, η περιοχή r_1 του DT_1 (βλέπε Εικόνα 5.2) μπορεί να περιγραφεί ως:

$$r.s = \{(10 \leq Age \leq 40) \wedge (20 \leq Income \leq 60)\}$$

$$r.m_D = [n_{c_1} = 20, n_{c_2} = 10]$$

Όπως φαίνεται από το παράδειγμα αυτό, τόσο το $r.s$ όσο και το $r.m_D$ μπορούν να εξαχθούν κατευθείαν από το δέντρο απόφασης, χωρίς επιπλέον επεξεργασία.

Πιθανότητα περιοχής, $\mathbf{P(r)}$: Η *πιθανότητα μιας περιοχής* (region probability) αναπαριστά την πιθανότητα κάποιο στιγμιότυπο x να ακολουθήσει το αντίστοιχειο μονοπάτι του δέντρου απόφασης. Τυπικά, αυτή η πιθανότητα δίνεται από τον τύπο:

$$P(r) \equiv Pr(x \in r) = \int_r P(A)dA$$

όπου $P(A)$ είναι η συνάρτηση πυκνότητας πιθανότητας των στιγμιοτύπων. Ωστόσο, καθώς δεν γνωρίζουμε την ακριβή μορφή της $P(A)$, θα πρέπει να χρησιμοποιήσουμε τα δεδομένα για να κάνουμε μία εκτίμηση αυτή της συνάρτησης.

Πιο συγκεκριμένα, αν θεωρήσουμε ένα σύνολο εκπαίδευσης D για τη δημιουργία του δέντρου απόφασης DT, μπορούμε να κάνουμε μία *εκτίμηση της $P(r)$ με βάση το σύνολο εκπαίδευσης* (training set dependent estimate) ως ακολούθως.

$$\mathbf{P_D(r)} = \frac{|r.m_D|}{N_D} \tag{5.1}$$

Αυτό είναι απλά το ποσοστό των στιγμιοτύπων εκπαίδευσης (N_D) που καταλήγουν στην r ($|r.m_D|$).

Υιοθετούμε το συμβολισμό P για την πραγματική κατανομή και τον συμβολισμό $\mathbf{P}$ για την εκτίμησή/ προσέγγισή της.

Το διάνυσμα:

$$\mathbf{P_D(A)} = [\mathbf{P_D(r_i)}|r_i \in R_{DT}] \tag{5.2}$$

αποτελεί μία προσέγγιση της $P(A)$, με βάση το σύνολο δεδομένων D. Μπορούμε να φανταστούμε αυτό το διάνυσμα σαν ένα ιστόγραμμα, όπου τα κουτιά είναι οι περιοχές του R_{DT}.

Απο κοινού πιθανότητα περιοχής - κλάσης, $\mathbf{P(r, c)}$: Εκτός της $P(A)$, μπορούμε να προσεγγίσουμε και την $P(A, C)$ εκμεταλευόμενοι τις ποσοτικές συνιστώσες των περιοχών, οι οποίες περιγράφουν την κατανομή των στιγμιοτύπων εκπαίδευσης στις διάφορες κλάσεις του προβλήματος.

Η *από κοινού πιθανότητα περιοχής - κλάσης* (joint region-class probability) για την εμφάνιση της περιοχής r με κλάση c δίνεται από τον τύπο:

$$\mathbf{P_D(r, c)} = \frac{r.n_c}{N_D} \tag{5.3}$$

Αυτό είναι απλά το ποσοστό των στιγμιοτύπων εκπαίδευσης (N_D) που καταλήγουν στην περιοχή r και ανήκουν στην κλάση c ($r.n_c$). Η εκτίμηση αυτή επίσης εξαρτάται από το σύνολο εκπαίδευσης D.

Το διάνυσμα:

$$\mathbf{P_D(r, C)} = [\mathbf{P_D(r, c_j)}, j = 1 \ldots k]$$

περιγράφει την απο κοινού κατανομή πιθανότητας περιοχής - κλάσης στην r. Μπορούμε να φανταστούμε το διάνυσμα αυτό σαν ένα ιστόγραμμα μίας διάστασης με k κουτιά, ένα για κάθε κλάση.

Ο πίνακας:

$$\mathbf{P_D}(\mathbf{A}, \mathbf{C}) = [\mathbf{P_D}(\mathbf{r_i}, \mathbf{c_j}) | r_i \in R_{DT}, c_j \in C] \tag{5.4}$$

όπου κάθε γραμμή αντιστοιχεί σε μία περιοχή $r_i \in R_{DT}$ και κάθε στήλη σε μία κλάση $c_j \in C$, αποτελεί μία προσέγγιση της από κοινού κατανομής $P(A, C)$ με βάση το σύνολο δεδομένων D. Η προσέγγιση αυτή έχει τη μορφή ενός πολυδιάστατου ιστογράμματος όπου τα κουτιά του ιστογράμματος ορίζονται από το $R_{DT} \times C$.

Υπο συνθήκη πιθανότητα της κλάσης με βάση την περιοχή, $\mathbf{P}(\mathbf{c}|\mathbf{r})$: Επιπλέον, η ποσοτική συνιστώσα μπορεί να μας δώσει μία προσέγγιση της υπο συνθήκη πιθανότητας της κλάσης c δοθείσας της περιοχής r:

$$\mathbf{P_D}(\mathbf{c}|\mathbf{r}) = \frac{r.n_c}{|r.m_D|}$$

Αυτό είναι απλά το ποσοστό των στιγμιοτύπων της περιοχής ($|r.m_D|$) που ανήκουν στην κλάση c ($r.n_c$).

Το διάνυσμα:

$$P_D(C|r) = [\mathbf{P_D}(\mathbf{c_j}|\mathbf{r}), j = 1 \ldots k]$$

περιγράφει την υπο συνθήκη κατανομή της κλάσης στην περιοχή r. Μπορούμε να φανταστούμε αυτό το διάνυσμα ως ένα μονοδιάστατο ιστόγραμμα με k κουτιά, ένα για καθεμία από τις κλάσεις του προβλήματος.

Ο πίνακας:

$$\mathbf{P_D}(\mathbf{C}|\mathbf{A}) = [\mathbf{P_D}(\mathbf{c_j}|\mathbf{r_i}) | r_i \in R_{DT}, c_j \in C] \tag{5.5}$$

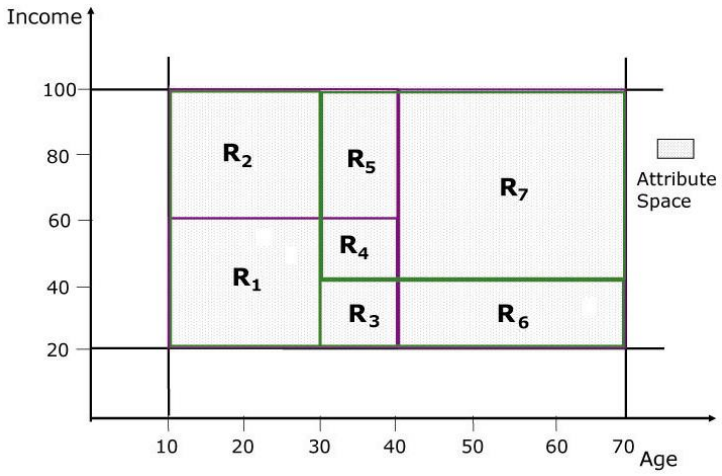
όπου κάθε γραμμή αντιστοιχεί σε μία περιοχή $r_i \in R_{DT}$ και κάθε στήλη αντιστοιχεί σε μία κλάση $c_j \in C$, αποτελεί μία προσέγγιση της $P(C|A)$ με βάση το σύνολο δεδομένων D. Η προσέγγιση αυτή έχει τη μορφή ενός πολυδιάστατου ιστογράμματος όπου τα κουτιά του ιστογράμματος ορίζονται από το $R_{DT} \times C$.

5.3.2 Επικάλυψη των τμηματοποιήσεων δέντρων απόφασης

Έστω R_{DT_1} και R_{DT_2} είναι οι τμηματοποιήσεις που ορίζουν τα δέντρα απόφασης DT_1 και DT_2, αντίστοιχα. Από την επικάλυψη των δύο τμηματοποίσεων προκύπτει μία πιο εκλεπτυσμένη τμηματοποίηση $R_{DT_1 \times DT_2}$, κάθε περιοχή r της οποίας είναι το αποτέλεσμα της επικάλυψης μιας περιοχής $r_i \in R_{DT_1}$ με κάποια περιοχή $r_j \in R_{DT_2}$, δηλαδή $r = r_i \cap r_j$. Στην Εικόνα 5.3 παρουσιάζουμε το αποτέλεσμα της επικάλυψης των τμηματοποίησεων των δέντρων απόφασης DT_1 και DT_2.

Θα πρέπει να υπσολογίσουμε την πιθανότητα της περιοχής $P(r)$, την από κοινού πιθανότητα της περιοχής και της κλάσης $P(r, c)$ και την υπο συνθήκη πιθανότητα $P(c|r)$ της κλάσης δοθείσας της περιοχής, για κάθε περιοχή $r \in R_{DT_1 \times DT_2}$ και κάθε κλάση $c \in C$. Για το σκοπό αυτό, βασιζόμαστε στην παρατήρηση πως κάθε περιοχή $r \in R_{DT_1 \times DT_2}$ αποτελεί επίσης ένα υπερ-τετράγωνο και μπορεί συνεπώς να περιγραφεί μέσω μίας δομικής και μιας ποσοτικής συνιστώσας.

Στην Εικόνα 5.4, παραθέτουμε το αποτέλεσμα της επικάλυψης των περιοχών $R_1 \in DT_1$ και $R_3 \in DT_2$.



Σχήμα 5.3: Η τμηματοποίηση $R_{DT_1 \times DT_2}$

5.3.2.1 Δομική συνιστώσα για τις περιοχές επικάλυψης

Η *δομική συνιστώσα* της περιοχής $r_i \cap r_j$ της επικάλυψης μπορεί εύκολα να οριστεί μέσω της τομής των περιοχών r_i, r_j που συμμετέχουν στη δημιουργία της:

$$r_i \cap r_j.s := \{\wedge t_i(a_i), i = 1 \dots m\}$$

$$t(a) := \min_a(r_i \cap r_j) \leq a \leq \max_a(r_i \cap r_j)$$

$$\min_a(r_i \cap r_j) := \max(\min_a(r_i), \min_a(r_j))$$

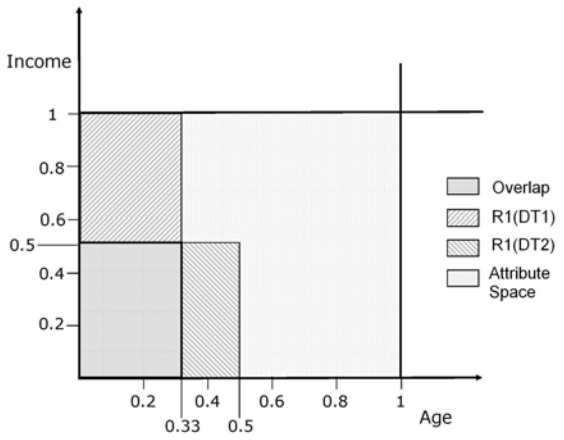
$$\max_a(r_i \cap r_j) := \min(\max_a(r_i), \max_a(r_j))$$

Αν $\max_a(r_i \cap r_j) \leq \min_a(r_i \cap r_j)$, η περιοχή επικάλυψης $r_i \cap r_j$ δεν ορίζεται καθώς οι περιοχές δεν τέμνονται.

5.3.2.2 Ποσοτική συνιστώσα για τις περιοχές επικάλυψης

Η εκτίμηση της ποσοτικής συνιστώσας της περιοχής επικάλυψης $r_i \cap r_j$ εξαρτάται από το σύνολο δεδομένων. Οι προφανείς επιλογές για το σύνολο δεδομένων είναι D_1, D_2 και $D_1 \cup D_2$. Ωστόσο, ακόμα και αν δεν έχουμε πλέον πρόσβαση στα σύνολα αυτά, θα μπορούσαμε να κάνουμε ακόμη μία εκτίμηση για την ποσοτική συνιστώσα των περιοχών της επικάλυψης με βάση τις ποσοτικές συνιστώσες των περιοχών των αρχικών τμηματοποιήσεων R_{DT_1} και R_{DT_2}. Αναφερόμαστε στην πρώτη περίπτωση ως εκτίμηση πιθανότητας με βάση τα δεδομένα και στη δεύτερη περίπτωση ως εκτίμηση πιθανότητας με βάση τα πρότυπα. Αναλύουμε στη συνέχεια και τις δύο περιπτώσεις.

- **Εκτίμηση πιθανότητας με βάση τα δεδομένα:** Αν έχουμε πρόσβαση στα πρωτογενή σύνολα δεδομένων, μπορούμε να υπολογίσουμε το *ακριβές μέτρο* των περιοχών επικάλυψης απλά προβάλλοντας κάθε σύνολο δεδομένων $D \in \{D_1, D_2, D_1 \cup D_2\}$ στο $R_{DT_1 \times DT_2}$.



Σχήμα 5.4: Η επικάλυψη των περιοχών $R_1 \in DT_1$ και $R_3 \in DT_2$ (οι τιμές των στιγμιοτύπων έχουν κανονικοποιηθεί στο διάστημα [0-1])

Έτσι, προβάλλοντας το D_1 στο $R_{DT_1 \times DT_2}$ μπορούμε να έχουμε το ακριβές μέτρο $r_i \cap r_j.m_{D_1}$:

$$r_i \cap r_j.m_{D_1} = \frac{|t \in r_i \cap r_j, t \in D_1|}{N_{D_1}} \tag{5.6}$$

Με παρόμοιο τρόπο, προβάλλοντας το D_2 στο $R_{DT_1 \times DT_2}$ μπορούμε να έχουμε το ακριβές μέτρο $r_i \cap r_j.m_{D_2}$:

$$r_i \cap r_j.m_{D_2} = \frac{|t \in r_i \cap r_j, t \in D_2|}{N_{D_2}} \tag{5.7}$$

Τέλος, προβάλλοντας το $D_1 \cup D_2$ στο $R_{DT_1 \times DT_2}$ μπορούμε να έχουμε το ακριβές μέτρο $r_i \cap r_j.m_{D_1 \cup D_2}$:

$$r_i \cap r_j.m_{D_1 \cup D_2} = \frac{|t \in r_i \cap r_j, t \in D_1 \cup D_2|}{N_{D_1} + N_{D_2}} \tag{5.8}$$

- **Εκτίμηση πιθανότητας με βάση τα πρότυπα:** Ακόμα και αν δεν έχουμε πρόσβαση στα πρωτογενή σύνολα δεδομένων (για λόγους π.χ. ιδιωτικότητας ή απόδοσης), μπορούμε να κάνουμε μία εκτίμηση του αναμενόμενου μέτρου για κάθε περιοχή επικάλυψης $r_i \cap r_j \in R_{DT_1 \times DT_2}$ χρησιμοποιώντας τις ποσοτικές συνιστώσες των αρχικών περιοχών $r_i \in R_{DT_1}$ και $r_j \in R_{DT_2}$. Το αναμενόμενο μέτρο της περιοχής επικάλυψης $r_i \cap r_j$ με βάση το σύνολο δεδομένων D_1 είναι:

$$r_i \cap r_j.m_{D_1} = r_i.m_{D_1} \frac{V(r_i \cap r_j)}{V(r_i)} \tag{5.9}$$

όπου ο όρος $\frac{V(r_i \cap r_j)}{V(r_i)}$ αναφέρεται στον όγκο που καταλαμβάνει η περιοχή επικάλυψης $r_i \cap r_j$ σε σχέση με τον όγκο που καταλαμβάνει η περιοχή r_i. Δεδομένου ότι οι περιοχές που ορίζει η τμηματοποίηση ενός δέντρου απόφασης είναι υπερ-τετράγωνα παράλληλα στους άξονες των γνωρισμάτων ισχύει ότι:

$$V(r) = \prod_{a_i} \frac{|t(a_i)|}{|dom(a_i)|}$$

Ο όρος $\frac{|t(a_i)|}{|dom(a_i)|}$ αναπαριστά τη σχετική σημαντικότητα του γνωρίσματος a_i στην περιοχή r. Αν υποθέσουμε μία ομοιόμορφη κατανομή $U(A)$ των στιγμιοτύπων πάνω στο χώρο γνωρισμάτων, τότε ισχύει ότι $V(r) = P(r)$. Στην Εξίσωση 5.9, ωστόσο, υιοθετούμε μία μεσαία υπόθεση. Συγκεκριμένα, υποθέτουμε ότι τα στιγμιότυπα του D_1 είναι ομοιόμορφα κατανεμημένα μέσα στην περιοχή r_i του R_{DT_1}, αντί της υπόθεσης ότι τα στιγμιότυπα είναι ομοιόμορφα κατανεμημένα μέσα σε όλο το χώρο των γνωρισμάτων.

Ακολουθώντας την ίδια λογική όπως στην Εξίσωση 5.9, το αναμενόμενο μέτρο της $r_i \cap r_j$ με βάση το σύνολο δεδομένων D_2 είναι:

$$r_i \cap r_j.m_{D_2} = r_j.m_{D_2} \frac{V(r_i \cap r_j)}{V(r_j)} \tag{5.10}$$

Να σημειώσουμε πως και σε αυτό τον υπολογισμό υποθέτουμε ότι τα στιγμιότυπα του D_2 είναι ομοιόμορφα κατανεμημένα μέσα στην περιοχή r_j που ορίζεται από την τμηματοποίηση του DT_2.

Τέλος, αν υποθέσουμε ότι τα δύο σύνολα δεδομένων προέρχονται από την ίδια κατανομή $P(A)$, μπορούμε να έχουμε μία εκτίμηση για το αναμενόμενο μέτρο της $r_i \cap r_j$ με βάση το σύνολο δεδομένων της ένωσης, $D_1 \cup D_2$:

$$r_i \cap r_j.m_{D_1 \cup D_2} \quad = \quad r_i \cap r_j.m_{D_1} + r_i \cap r_j.m_{D_2}$$

Μέχρι στιγμής, είδαμε πως μπορούμε να εκτιμήσουμε την πιθανότητα κάθε περιοχής στην επικάλυψη των τμηματοποιήσεων ανάλογα με το αν έχουμε πρόσβαση στα πρωτογενή δεδομένα ή όχι. Όπως και στην περίπτωση της απλής τμηματοποίησης ενός δέντρου απόφασης (βλέπε Ενότητα 5.3.1), μπορούμε να χρησιμοποιήσουμε αυτές τις εκτιμήσεις για να προσεγγίσουμε τις κατανομές $P(A)$, $P(A, C)$, $P(C|A)$. Ανάλογα με ποιο σύνολο δεδομένων, $D \in \{D_1, D_2, D_1 \cup D_2\}$, χρησιμοποιούμε για να υπολογίσουμε τις ποσοτικές συνιστώσες των περιοχών επικάλυψης, παίρνουμε τις αντίστοιχες εκτιμήσεις των $\mathbf{P_D(A)}$, $\mathbf{P_D(A, C)}$ και $\mathbf{P_D(C|A)}$ με βάση την τμηματοποίηση επικάλυψης $R_{DT_1 \times DT_2}$. Για να ξεχωρίσουμε την περίπτωση που τα μέτρα υπολογίζονται προσπελαύοντας τα πρωτογενή σύνολα δεδομένων (Εξισώσεις 5.6, 5.7) από την περίπτωση που τα μέτρα υπολογίζονται με βάση την υπόθεση της ομοιόμορφης κατανομής των στιγμιοτύπων μέσα σε μία περιοχή (Εξισώσεις 5.9, 5.10), χρησιμοποιούμε τους δείκτες ^{Q} και ^{U}, αντίστοιχα.

5.3.3 Μέτρα απόστασης για δέντρα απόφασης και σύνολα δεδομένων κατηγοριοποίησης

Στην προηγούμενη ενότητα, περιγράψαμε μεθόδους για την εκτίμηση των κατανομών $\mathbf{P_D(A)}$, $\mathbf{P_D(A, C)}$ και $\mathbf{P_D(C|A)}$ με βάση την τμηματοποίηση $R_{DT_1 \times DT_2}$ και για

τα διάφορα σύνολα δεδομένων $D \in \{D_1, D_2, D_1 \cup D_2\}$. Οι εκτιμήσεις αυτές μπορούν να χρησιμοποιηθούν για τον υπολογισμό της ομοιότητας μεταξύ δέντρων απόφασης και συνόλων δεδομένων κατηγοριοποίησης.

Πριν προχωρήσουμε με τον ορισμό των μέτρων ομοιότητας, παρουσιάζουμε μία συνάρτηση ομοιότητας μεταξύ ιστογραμμάτων, καθώς όλες μας οι εκτιμήσεις έρχονται με τη μορφή ιστογραμμάτων. Έστω P, Q είναι δύο εκτιμήσεις πυκνότητας πιθανότητας για μία τυχαία μεταβλητή X που προέρχονται από δύο διαφορετικούς πληθυσμούς. Έστω επίσης ότι τα ιστογράμματα που αντιστοιχούν στις P, Q έχουν την ίδια δομή (δηλαδή αποτελούνται από τα ίδια κουτιά (bins».

Ο *συντελεστής συγγένειας* (affinity coefficients) μεταξύ των P και Q ορίζεται ως εξής:

$$s(P, Q) = \sum_i \sqrt{P_i Q_i}$$

Το σκορ που προκύπτει ανήκει στο εύρος $[0 - 1]$.

Με βάση τους συντελεστή συγγένειας που μόλις ορίσαμε και τις διαφορετικές εκτιμήσεις για τα στατιστικά των περιοχών επικάλυψης που παρουσιάσαμε σε προηγούμενες ενότητες αυτού του κεφαλαίου, μπορούμε να ορίσουμε διάφορα μέτρα ομοιότητας μεταξύ δέντρων απόφασης και συνόλων δεδομένων κατηγοριοποίησης:

Περίπτωση α: Μπορούμε να υπολογίσουμε την *ομοιότητα δύο συνόλων δεδομένων* D_1, D_2 με βάση τις κατανομές πιθανότητας των δύο αυτών συνόλων πάνω στο χώρο των γνωρισμάτων του προβλήματος, δηλαδή με βάση τα $P_{D_1}(A)$, $P_{D_2}(A)$, αντίστοιχα. Αυτό μπορεί να γίνει κατευθείαν με βάση τη σχέση:

$$s(\mathbf{P_{D_1}}(\mathbf{A}), \mathbf{P_{D_2}}(\mathbf{A})) \tag{5.11}$$

Αυτό το μέτρο ομοιότητας μπορεί να χρησιμοποιηθεί για να κρίνει αν δύο σύνολα δεδομένων προέρχονται από την ίδια κατανομή $P(A)$. Ανάλογα με το αν έχουμε πρόσβαση στα πρωτογενή σύνολα δεδομένων, οι εκτιμήσεις $\mathbf{P_{D_i}}(\mathbf{A}), i = \{1, 2\}$ μπορεί να είναι οι $\mathbf{P^Q_{D_i}}(\mathbf{A})$ ή οι $\mathbf{P^U_{D_i}}(\mathbf{A})$.

Περίπτωση β: Μπορούμε να υπολογίσουμε την *ομοιότητα δύο δέντρων απόφασης* DT_1, DT_2 με βάση τις προβλέψεις τους. Αυτό είναι ένα μέτρο *σημασιολογικής ομοιότητας* μεταξύ δύο δέντρων απόφασης, δηλαδή, υπολογίζει πόσο όμοιες είναι οι έννοιες που περιγράφουν τα δύο δέντρα. Η ομοιότητα αυτή αντιστοιχεί στο ποσοστό των περιπτώσεων που τα δύο δέντρα παράγουν την ίδια πρόβλεψη για στιγμιότυπα που προέρχονται από μία συγκεκριμένη κατανομή πάνω στο χώρο των γνωρισμάτων.

Ορίζουμε πρώτα το διάνυσμα:

$$\mathbf{I}(\mathbf{C}|\mathbf{A}) = [I(r_i.cl, r_j.cl)|r_i \cap r_j \in R_{DT_1 \times DT_2}]$$

το οποίο συμβολίζει κατά πόσο δύο δέντρα απόφασης συμφωνούν ή διαφωνούν στις προβλέψεις τους σχετικά με τις περιοχές που προκύπτουν από την επικάλυψη $R_{DT_1 \times DT_2}$. Η $I(r_i.cl, r_j.cl)$ επιστρέφει 1 αν τα δύο δέντρα παράγουν την ίδια πρόβλεψη σχετικά με την περιοχή $r_i \cap r_j$, δηλαδή, αν $r_i.cl = r_j.cl$, διαφορετικά επιστρέφει 0. Το εσωτερικό γινόμενο:

$$S(DT_1, DT_2) = \mathbf{I}(\mathbf{C}|\mathbf{A})'\mathbf{P}(\mathbf{A}) \tag{5.12}$$

υπολογίζει την ομοιότητα στις προβλέψεις των DT_1, DT_2 σε σχέση με την κατανομή $P(A)$. Η ομοιότητα ισούται με το άθροισμα των πιθανοτήτων των περιοχών $r_i \cap r_j$ για τις οποίες υπάρχει συμφωνία των δύο δέντρων απόφασης ως προς την κλάση. Στην παραπάνω εξίσωση, ο συμβολισμός X' συμβολίζει τον αντίστροφο του πίνακα X.

Ένα θέμα που προκύπτει εδώ είναι ποια εκτίμηση της $P(A)$ πρέπει να χρησιμοποιήσουμε. Οι πιθανές επιλογές είναι οι επόμενες:

- *η ομοιόμορφη κατανομή,* $U(A)$. Στην περίπτωση αυτή, η συμφωνία θα εξεταστεί πάνω σε όλους τους πιθανούς κόσμους εισόδου. Με βάση την υπόθεση αυτή, η πιθανότητας μιας περιοχής $r_i \cup r_j$ δίνεται από τον όγκο της. Συνεπώς, η ομοιότητα μεταξύ δύο δέντρων απόφασης ισούται με τον συνολικό όγκο των περιοχών στις οποίες τα δέντρα απόφασης συμφωνούν ως προς την κλάση που προβλέπουν. Στην περίπτωση αυτή, η Εξίσωση 5.12 αντιστοιχεί στη σημασιολογική ομοιότητα μεταξύ δύο δέντρων απόφασης έτσι όπως ορίστηκε από τον Turney [78]. Να σημειώσουμε ωστόσο πως για τον υπολογισμό αυτό εμείς *δεν απαιτούμε* την κατασκευή ενός τεχνητού συνόλου δεδομένων ελέγχου προερχόμενου από την $U(A)$, όπως στην [78].
- *μία εξαρτόμενη από το σύνολο δεδομένων κατανομή* $\mathbf{P_D}(\mathbf{A})$, όπου το D θα μπορούσε να είναι ένα από τα D_1, D_2 και $D_1 \cup D2$ σύνολα δεδομένων. Στην περίπτωση αυτή, υποθέτουμε ότι τα στιγμιότυπα ακολουθούν την κατανομή του συνόλου δεδομένων $D \in \{D_1, D_2, D_1 \cup D_2\}$. Η ένωση, $D_1 \cup D2$, είναι η πιο κατάλληλη επιλογή αν τα δέντρα έχουν εξαχθεί από σύνολα δεδομένω που προέρχονται από την ίδια κατανομή και εμείς ενδιαφερόμαστε να υπολογίσουμε την ομοιότητά τους σε σχεση με τη συγκεκριμένη κατανομή.
- τέλος, η $P(A)$ θα μπορούσε να είναι μία κατανομή *διαφορετική* από τις κατανομές από τις οποίες προέρχονται τα σύνολα εκπαίδευσης.

Περίπτωση γ: Μπορούμε επίσης να υπολογίσουμε την ομοιότητα των δύο συνόλων δεδομένων με βάση την υπο-συνθήκη κατανομή $P(C|A)$ που τα δέντρα απόφασης, τα οποία εξάγονται από τα συγκεκριμένα σύνολα δεδομένων, ορίζουν πάνω στο χώρο γνωρισμάτων. Ωε φιρστ δεφινε τηε εχτορ:

$$\mathbf{S}(\mathbf{C}|\mathbf{A}) = [s(\mathbf{P_{D_1}}(\mathbf{C}|\mathbf{A})[r_i,], \mathbf{P_{D_2}}(\mathbf{C}|\mathbf{A})[r_j,])| \; r_i \cap r_j \in R_{DT_1 \times DT_2}]$$

Η $\mathbf{S}(\mathbf{C}|\mathbf{A})$ έχει την ίδια δομή με την $\mathbf{I}(\mathbf{C}|\mathbf{A})$, αλλά η δυαδική $(0/1)$ συνάρτηση ομοιότητας $I(.,.)$ έχει αντικατασταθεί με την συνάρτηση $s(.,.)$, η οποία υπολογίζει την ομοιότητα των υπο συνθήκη κατανομών των D_1 ανδ D_2 στην περιοχή $r_i \cap r_j$. Το εσωτερικό γινόμενο:

$$S(D_1, D_2) = \mathbf{S}(\mathbf{C}|\mathbf{A})'\mathbf{P}(\mathbf{A}) \tag{5.13}$$

παρέχει ένα μέτρο της ομοιότητας των δύο συνόλων δεδομένων με βάση τις υπο-συνθήκη κατανομές τους πάνω σε ένα χώρο γνωρισμάτων που ακολουθεί την κατανομή $P(A)$.

Περίπτωση δ: Τέλος, μπορούμε να υπολογίσουμε την ομοιότητα μεταξύ των κατανομών από κοινού πιθανότητας των γνωρισμάτων και της κλάσης, $P_{D_1}(A, C)$ και $P_{D_2}(A, C)$, εφαρμόζοντας απλά τον συντελεστή συγγένειας ως ακολούθως:

$$s(\mathbf{P_{D_1}}(\mathbf{A}, \mathbf{C}), \mathbf{P_{D_2}}(\mathbf{A}, \mathbf{C})) \tag{5.14}$$

Η $\mathbf{P_{D_i}}(\mathbf{A}, \mathbf{C})$ είναι η εκτίμηση της $P_{D_i}(A, C)$ με βάση την επικάλυψη. Αν δύο σύνολα δεδομένων προέρχονται από την ίδια κατανομή $P(A)$, τότε το μέτρο αυτό είναι ισοδύναμο με το $S(D_1, D_2)$ της Εξίσωσης 5.13. Αυτή είναι η προσέγγιση που προτείνει το FOCUS [25] για τον υπολογισμό της απόστασης δύο συνόλων δεδομένων. Η διαφορά έγκειται στο γεγονός ότι αντί του συντελεστή συγγένειας το FOCUS χρησιμοποιεί μία συνάρτηση διαφοράς f (π.χ. απόλυτη ή σχετική διαφορά) για να υπολογίσει την ομοιότητα μέσα σε κάθε περιοχή και μία συνάρτηση συνάθροισης g (π.χ. sum ή max) για να αθροίσει τα σκορ των περιοχών επικάλυψης σε ένα συνολικό σκορ.

Στην ενότητα αυτή, παροσιάσαμε ένα γενικό πλαίσιο για τη σύγκριση δέντρων απόφασης και συνόλων δεδομένων κατηγοριοποίησης με βάση τα μοντέλα δέντρων απόφασης. Βάσει αυτού του πλαισίου, μπορούμε να υπολογίσουμε την ομοιότητα δύο συνόλων δεδομένων σε σχέση με διάφορες κατανομές πιθανότητας: ι) την κατανομή του χώρου γνωρισμάτων $P(A)$ (Εξίσωση 5.11), ιι) την από κοινού κατανομή γνωρισμάτων και κλάσης $P(A, C)$ (Εξίσωση 5.14) και ιιι) την υπόσυνθήκη κατανομή της κλάσης με βάση τα γνωρίσματα $P(C|A)$ (Εξίσωση 5.13). Μπορούμε επίσης να χρησιμοποιήσουμε αυτό το πλαίσιο για να υπολογίσουμε την σημασιολογική ομοιότητα μεταξύ δέντρων απόφασης (Εξίσωση 5.12) κάτω από διαφορετικές υποθέσεις σε σχέση με την κατανομή πυκνότητας πιθανότητας του χώρου γνωρισμάτων. Αυτή την τελευταία κατεύθυνση θα εξερευνήσουμε περαιτέρων στην επόμενη ενότητα, μέσω πειραμάτων.

5.4 Πειραματική αξιολόγηση

Η σημασιολογική ομοιότητα μεταξύ δύο μοντέλων κατηγοριοποίησης M_1, M_2 ορίζεται ως το ποσοστό των περιπτώσεων όπου τα δύο μοντέλα παράγουν τις ίδιες προβλέψεις σχετικά με την κλάση στιγμιοτύπων που προέρχονται από μία κατανομή $P(A)$ πάνω στο χώρο γνωρισμάτων. Ο Turney [78] όρισε ένα μέτρο σημασιολογικής ομοιότητας για μοντέλα κατηγοριοποίησης, καλούμενο *συμφωνία* (agreement), ως την πιθανότητα τα δύο μοντέλα να παράξουν τις ίδιες προβλέψεις πάνω σε όλα τα πιθανά στιγμιότυπα που προέρχονται από την ομοιόμορφη κατανομή $U(A)$ πάνω στο χώρο γνωρισμάτων. Ο Turney υπολογίζει εμπερικά την συμφωνία μεταξύ δύο μοντέλων κατηγοριοποίησης, εφαρμόζοντας και τα δύο μοντέλα σε ένα σύνολο δεδομένων ελέγχου D_H το οποίο αποτελείται από στιγμιότυπα που προέρχονται από την κατανομή $U(A)$ και ελέγχοντας εν συνεχεία το ποσοστό των περιπτώσεων που τα μοντέλα παράγουν τις ίδιες προβλέψεις. Το επιχείρημα σχετικά με τη χρήση της ομοιόμορφης κατανομής $U(A)$ αντί της κατανομής $P(A)$ από την οποία προέρχονται τα δεδομένα είναι ότι η συμφωνία μεταξύ δύο εννοιών πρέπει να εξεταστεί σε όλους τους πιθανούς κόσμους εισόδου.

Εμείς πιστεύουμε πως σε μία πραγματική εφαρμογή, αυτό που ενδιαφέρει δεν είναι η ομοιότητα των δέντρων σε όλους τους πιθανούς κόσμους αλλά η ομοιότητά τους στον κόσμο από τον οποίο προέρχονται τα δεδομένα από τα οποία κατασκευάστηκαν

τα δέντρα. Έτσι, σε αντίθεση με την [78], προκειμένου να υπολογίσουμε τη σημασιολογική ομοιότητα δύο δέντρων απόφασης, κατασκευάζουμε το σύνολο δεδομένων D_H από την κατανομή $P(A)$ που "κυβερνά" το χώρο των γνωρισμάτων. Συμβολίζουμε με $S_H(DT_1, DT_2)$ τη σημασιολογική ομοιότητα μεταξύ των DT_1 και DT_2 - αυτή η ομοιότητα υπολογίζεται εμπειρικά στο σύνολο δεδομένων D_H εφαρμόζοντας τα δύο δέντρα απόφασης στο D_H και υπολογίζοντας το πλήθος των περιπτώσεων που τα δέντρα παράγουν τις ίδιες προβλέψεις. Η $S_H(DT_1, DT_2)$ αποτελεί την *πραγματική ομοιότητα* (ground truth) με την οποία θα συγκρίνουμε το μέτρο σημασιολογικής ομοιότητας που προτείνουμε.

5.4.1 Σχεδιασμός των πειραμάτων

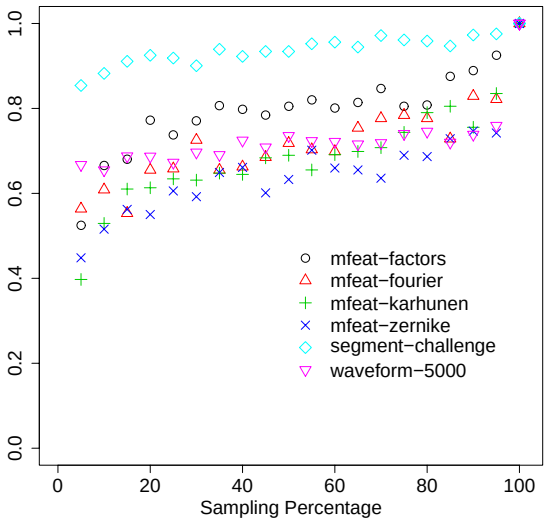
Καταρχήν χρειαζόμαστε έναν τρόπο να δημιουργούμε δέντρα απόφασης με διαφορετικό βαθμό σημασιολογικής ομοιότητας. Για το σκοπό αυτό, χωρίζουμε τυχαία ένα δοθέν σύνολο δεδομένων D σε δύο υποσύνολα, ένα α *σύνολο εκπαίδευσης* D_T που χρησιμοποιείται για την κατασκευή του δέντρου, και ένα *σύνολο ελέγχου* D_H που παίζει το ρόλο του hold out συνόλου για τον υπολογισμό της S_H ($|D_H| = \frac{1}{3}|D|$). Στη συνέχεια, δημιουργούμε από το D_T τυχαία υποσύνολα μεγέθους p ($p = 5\% \ldots 95\%$) με βήμα 5%. Σε κάθε υποσύνολο DT_p, εκπαιδεύουμε ένα δέντρο απόφασης και το συγκρίνουμε με το δέντρο απόφασης που εκπαιδεύτηκε στο συνολικό σύνολο εκπαίδευσης, DT_{100}. Στη συνέχεια, υπολογίζουμε τη σημασιολογική ομοιότητα μεταξύ του δέντρου απόφασης που προέρχεται από όλο το σύνολο εκπαίδευσης και του δέντρου απόφασης που προέρχεται από το δείγμα, δηλαδή, $S_H(DT_p, DT_{100})$, στο hold out σύνολο D_H.

Πειραματιστήκαμε με έξι διαφερετικά σύνολα δεδομένων, μία σύντομη περιγραφή των οποίων υπάρχει στον Πίνακα 5.2. Τα διαφορετικά *mfeat* σύνολα προέρχονται από το ίδιο πρόβλημα αναγνώρισης προτύπων που αποσκοπεί στην κατηγοριοποίηση αριθμών γραμμένων με το χέρι. Τα διάφορα σύνολα αντιστοιχούν στα διαφορετικά γνωρίσματα που χρησιμοποιούνται για την περιγραφή των αριθμών. Το *Waveform-5000* είναι ένα τεχνητό σύνολο δεδομένων όπου οι κλάσεις αντιστοιχούν στους διάφορους τύπους κυμάτων [12]. Στο σύνολο δεδομένων *segment-challenge*, [11], τα γνωρίσματα είναι υψηλού επιπέδου περιγραφές των περιοχών των εικόνων και ο σκοπός είναι η κατηγοριοποίηση κάθε εικόνας στη σωστή κλάση, π.χ., ουρανός, γρασίδι.

Πρώτα από όλα, πρέπει να βεβαιωθούμε ότι η διαδικασία που ακολουθήσαμε για τη δημιουργία των διαφόρων δέντρων απόφασης DT_p οδηγεί πραγματικά σε δέντρα απόφασης που εμφανίζουν διαφορετικά επίπεδα σημασιολογικής ομοιότητας σε σχέση με το DT_{100}. Αναμένουμε η τιμή της $S_H(DT_p, DT_{100})$ να αυξάνεται καθώς το p αυξάνεται και προσεγγίζει το 100%, καθώς το σύνολο εκπαίδευσης D_p που χρησιμοποιείται για την κατασκευή του DT_p γίνεται όλο και πιο όμοιο με το σύνολο εκπαίδευσης D_{100} που χρησιμοποιείται για την κατασκευή του DT_{100}. Αυτό όντως ισχύει όπως φαίνεται και στην Εικόνα 5.5, όπου αναπαριστούμε την S_H σε συνδυσμό με το μέγεθος του δείγματος p. Πράγματι υπάρχει μία ομαλή αύξηση στις τιμές της S_H καθώς το p αυξάνεται και πλησιάζει το 100%. Στόχος των πειραμάτων μας σε αυτή την ενότητα είναι να δούμε πως τα διάφορα μέτρα σημασιολογικής ομοιότητας που προτείνουμε συσχετίζονται με την εμπειρική σημασιολογική ομοιότητα S_H.

Σύνολο δεδομένων	Αρ. στιγμιοτύπων	Αρ. γνωρισμάτων	Αρ. κλάσεων
mfeat-factors	2,000	21	10
mfeat-fourier	2,000	76	10
mfeat-karhunen	2,000	64	10
mfeat-zernike	2,000	47	10
segment-challenge	2310	19	7
waveform-5000	5,000	40	3

Πίνακας 5.2: Περιγραφή των συνόλων δεδομένων



Σχήμα 5.5: Εξέλιξη της $S_H(DT_p, DT_100)$ με το μέγεθος του δείγματος p

5.4.2 Ποιοτική ανάλυση της προτεινόμενης σημασιολογικής ομοιότητας

Το μέτρο σημασιολογικής ομοιότητας δέντρων απόστασης $S(DT_1, DT_2)$ που προτείνουμε (Εξίσωση 5.12) εξαρτάται από την εκτίμηση της κατανομής $P(A)$ που "κυβερνά" το χώρο των γνωρισμάτων. Μάλιστα, ο υπολογισμός της ομοιότητας έχει νόημα για ένα συγκεκριμένο κόσμο, ο χώρος γνωρισμάτων του οποίου ακολουθεί μία συγκεκριμένη κατανομή $P(A)$. Στην περίπτωση αυτή, η ομοιότητα $S(DT_1, DT_2)$ είναι απλά το άθροισμα των τιμών πυκνότητας πιθανότητας των περιοχών $r_i \cap r_j$ στις οποίες τα δύο δέντρα συμφωνούν ως προς τις προβλέψεις τους. Όπως αναφέραμε ήδη, υπό την υπόθεση της ομοιόμορφης κατανομής των στιγμιοτύπων στο χώρο των γνωρισμάτων ($U(A)$), το αθροισμα αυτό ισούται με το άθροισμα των όγκων των περιοχών $r_i \cap r_j$. Επίσης, υπό την ίδια υπόθεση το $S(DT_1, DT_2)$ παρέχει τη σημασιολογική ομοιότητα του Turney [78] *χωρίς* μάλιστα να χρειάζεται να εφαρμόσουμε τα δέντρα στο hold out σύνολο. Δε θα ασχοληθούμε περαιτέρω με την υπόθεση της ομοιόμορφης κατανομής $U(A)$ ως τρόπου εκτίμησης της $P(A)$. Αντί αυτού, θα πειραματιστούμε με τρεις διαφορετικές αρ-

χικοποιήσεις της $S(DT_1, DT2)$, οι οποίες διαφοροποιούνται ως προς το ποια εκτίμηση της $P(A)$ χρησιμοποιούν. Συγκεκριμένα, θα διευρευνήσουμε περαιτέρω τις ακόλουες εκτιμήσεις της $P(A)$:

- $P^U_{D_1 \cup D_2}$: αυτή είναι η εκτίμηση της $P(A)$ που παίρνουμε όταν οι ποσοτικές συνιστώσες υπολογίζονται με βάση της υπόθεση της ομοιόμορφης κατανομής των στιγμιοτύπων στις περιοχές της τμηματοποίησης που ορίζει ένα δέντρο απόφασης, όπως στις Εξισώσεις 5.9, 5.10.

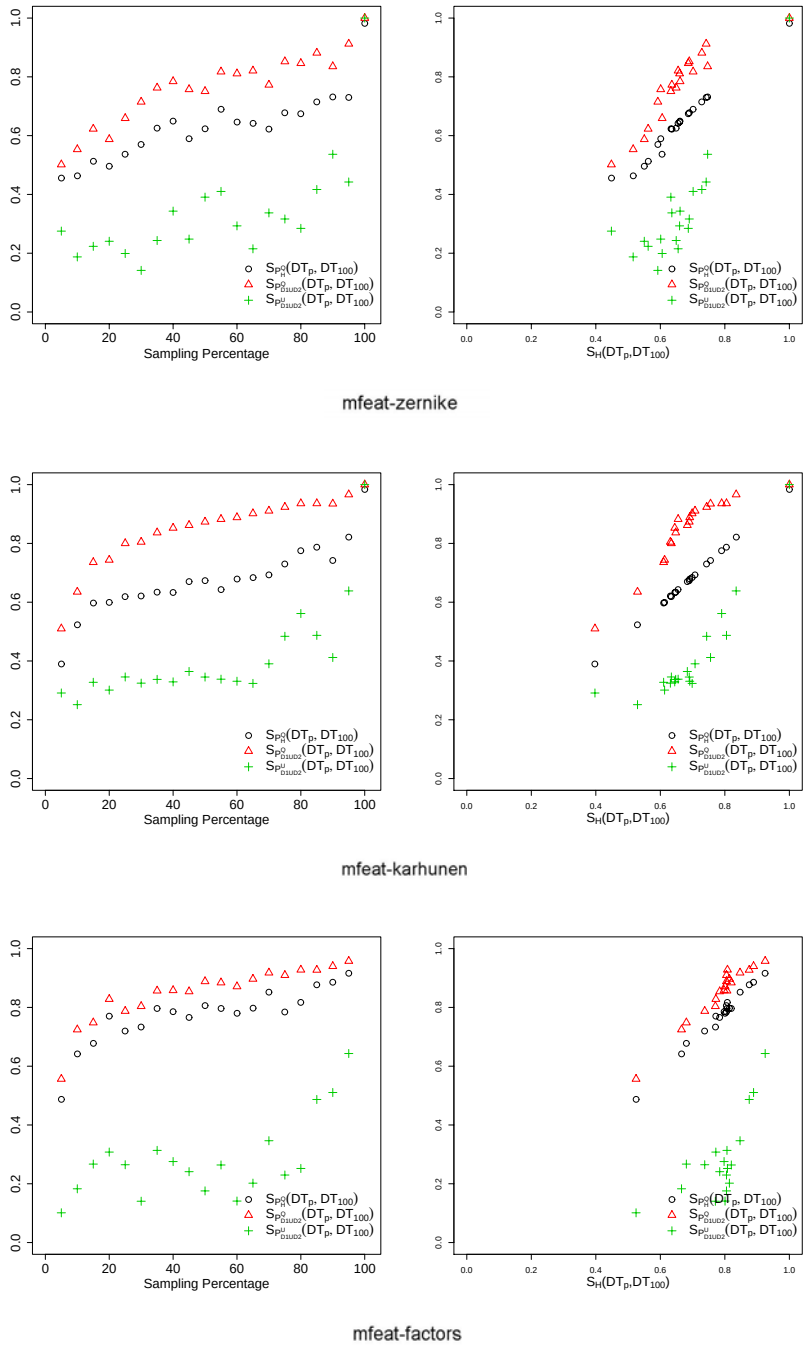
- $P^Q_{D_1 \cup D_2}$: αυτή είναι η εκτίμηση της $P(A)$ που παίρνουμε όταν οι ποσοτικές συνιστώσες υπολογίζονται προβάλλοντας κατευθείαν τα σύνολα δεδομένων D_p και D_{100} πάνω στην τμηματοποίηση της επικάλυψης $R_{DT_p \times DT_{100}}$.

- P^Q_H: αυτή είναι η εκτίμηση της $P(A)$ που παίρνουμε όταν το hold out σύνολο ελέγχου D_H προβάλλεται κατευθείαν πάνω στην τμηματοποίηση της επικάλυψης $R_{DT_p \times DT_{100}}$.

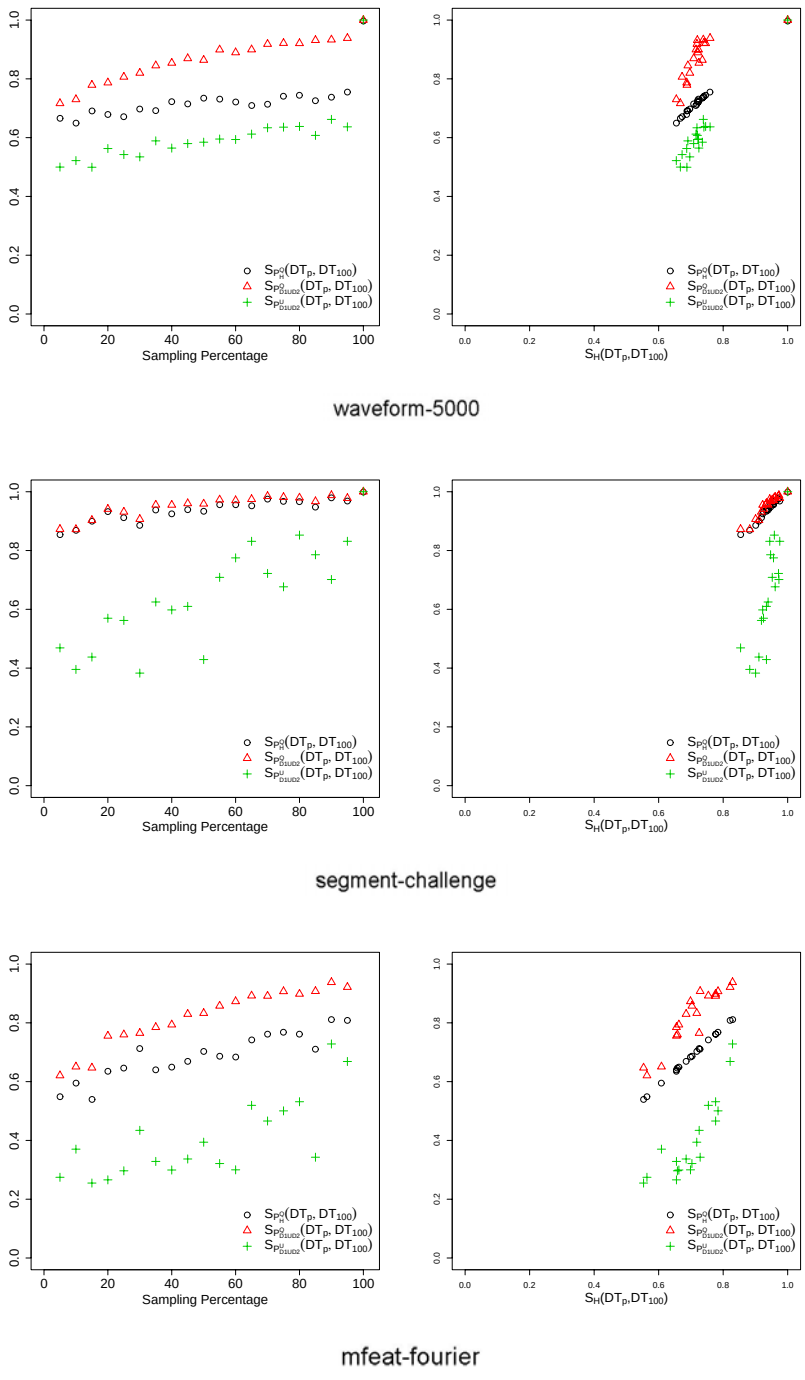
Κάθε μία από αυτές τις εκτιμήσεις P^Y_X της $P(A)$ οδηγεί σε μία διαφορετική αρχικοποίηση της $S(DT_1, DT_2)$, την οποία συμβολίζουμε με $S_{P^Y_X}(DT_1, DT_2)$. Να σημειώσουμε πως η σειρά με την οποία παρουσιάζονται οι διαφορετικές εκτιμήσεις P^Y_X εκφράζει μία αυξανόμενη γνώση σχετικά με την κατανομή $P(A)$.Η $P^U_{D_1 \cup D_2}$ υποθέτει τη λιγότερη γνώση σχετικά με την $P(A)$, καθώς, για να υπολογίσει τις ποσοτικές συνιστώσες των περιοχών της τμηματοποίησης της επικάλυψης, βασίζεται αποκλειστικά στις πληροφορίες που περιέχουν οι επιμέρους τμηματοποιήσεις των δέντρων απόφασης και στην υπόθεση της ομοιόμορφης κατανομής των στιγμιοτύπων στις περιοχές της τμηματοποίησης που ορίζει ένα δέντρο απόφασης πάνω στο χώρο των γνωρισμάτων. Η $P^Q_{D_1 \cup D_2}$ απαιτεί την εκτέλεση *επερωτήσεων* πάνω στα σύνολα δεδομένων D_1 και D_2 για να υπολογίσει τις ποσοτικές συνιστώσες των περιοχών της τμηματοποίησης της επικάλυψης. Συνεπώς, η εκτίμηση της $P(A)$ που παρέχει είναι πιο ακριβής σε σχέση με αυτή παρέχεται από την $P^U_{D_1 \cup D_2}$. Τέλος, η P^Q_H έχει πλήρη γνώση της $P(A)$, όπως αυτή η γνώση υπάρχει στο σύνολο δεδομένων D_H, καθώς αυτή η εκτίμηση προέρχεται εκτελώντας *επερωτήσεις* πάνω στο D_H. Συνεπώς, η $S_{P^Q_H}(DT_1, DT_2)$ θα πρέπει να συσχετίζεται τέλεια με την S_H.

Στην πρώτη στήλη των Εικόνων 5.6, 5.7 δείχνουμε πως μεταβάλλεται κάθε $S_{P^Y_X}(DT_p, DT_{100})$ σε σχέση με το μέγεθος του δείγματος p. Όλα τα μέτρα εμφανίζουν μία παρόμοια συμπεριφορά: η ομοιότητα αυξάνεται καθώς το p αυξάνεται. Επιπλέον, οι $S_{P^Q_{D_1 \cup D_2}}$ και $S_{P^Q_H}$ έχουν μία ομαλή συμπεριφορά, με μία σχεδόν σταθερή αύξηση των τιμών και μικρές διακυμάνσεις. Στην περίπτωση της $S_{P^U_{D_1 \cup D_2}}$, η ομοιότητα επίσης αυξάνεται καθώς το p αυξάνεται αλλά οι διακυμάνσεις εδώ είναι πολύ μεγαλύτερες (βλέπε τα σύνολα δεδομένων *mfeat-zernike*, *mfeat-factors*, *segment-challenge*, *mfeat-fourier*. Η $S_{P^Q_{D_1 \cup D_2}}$ συνεχώς υπερεκτιμά την ομοιότητα σε σχέση με την $S_{P^Q_H}$, ενώ η $S_{P^U_{D_1 \cup D_2}}$ την υποτιμά σημαντικά (να τονίσουμε πάλι εδώ πως η $S_{P^Q_H}$ ανταναλκλά την ιδανική συμπεριφορά).

Στη δεύτερη στήλη των Εικόνων 5.6, 5.7, βλέπουμε πως οι τρεις διαφορετικές εκδόσεις της $S_{P^Y_X}(DT_p, DT_{100})$ συσχετίζονται με την ομοιότητα-στόχο $S_H(DT_p, DT_{100})$. Όπως αναμέναμε, η $S_{P^Q_H}$ συσχετίζεται τέλεια με την S_H καθώς η εκτίμηση που χρησιμοποιεί για την $P(A)$ προέρχται από το σύνολο δεδομένων D_H στο οποίο έχει υπολογιστεί η $S_H(DT_p, DT_{100})$. Η $S_{P^Q_{D_1 \cup D_2}}$ υποτιμά συνεχώς την $S_H(DT_p, DT_{100})$,



Σχήμα 5.6: Εξέλιξη της σημασιολογικής ομοιότητας με το μέγεθος του δείγματος (πρώτη στήλη) και την S_H (δεύτερη στήλη)



Σχήμα 5.7: Εξέλιξη της σημασιολογικής ομοιότητας με το μέγεθος του δείγματος (πρώτη στήλη) και την S_H (δεύτερη στήλη)

ενώ η $S_{P^U_{D_1 \cup D_2}}$ την υπερεκτιμά σημαντικά. Η απόδοση της $S_{P^Q_{D_1 \cup D_2}}$ είναι αρκετά κοντά στην ιδανική απόδοση της $S_{P^Q_H}$, ειδικά στην περίπτωση των συνόλων δεδομένω *segment-challenge* και *mfeat-factors*, ενώ για το σύνολο δεδομένων *mfeat-karhunen* εμφανίζεται η μεγαλύτερη ασυμφωνία. Να σημειώσουμε εδώ πως τα σύνολα δεδομένων D_p, D_T και D_H προέρχονται όλα από την ίδια κατανομή $P(A)$. Η ασυμφωνία ανάμεσα στη συμπεριφορά της $S_{P^Q_{D_1 \cup D_2}}$ και της $S_{P^Q_H}$ μπορεί να οφείλεται στη δειγματοληψία. Καθώς το πλήθος των στιγμιοτύπων αυξάνεται, οι συμπεριφορές των $S_{P^Q_{D_1 \cup D_2}}$ και $S_{P^Q_H}$ θα συγκλίνουν καθώς οι εκτιμήσεις της $P(A)$ που χρησιμοποιούν οι δύο μέθοδοι για συγκλίνουν επίσης. Από την άλλη, η συμπεριφορά της $S_{P^U_{D_1 \cup D_2}}$ θα είναι παρόμοια με αυτή της $S_{P^Q_H}$ μόνο εφόσον η υπόθεση της ομοιόμορφης κατανομής μέσα στα όρια μιας περιοχής είναι έγκυρη για την κατανομή $P(A)$ που "κυβερνά" το D_H. Ωστόσο, όπως φαίνεται από τα πειραματικά αποτελέσματα για τα σύνολα δεδομένων που χρησιμοποιήσαμε εδώ, η υπόθεση αυτή δεν ισχύει.

5.4.3 Ποσοτική ανάλυση της προτεινόμενης σημασιολογικής ομοιότητας

Προκειμένου να ποσοτικοποιήσουμε την συμπεριφορά κάθε $S_{P^Y_X}(DT_p, DT_{100})$ υπολογίσαμε τους συντελεστές συσχέτισης Pearson (Pearson correlation coefficients) με την $S_H(DT_p, DT_{100})$. Τα αποτελέσματα παρατίθενται στον Πίνακα 5.3, όπου φαίνεται πως η $S_{P^Q_{D_1 \cup D_2}}$ εμφανίζει μία πολύ ισχυρή συσχέτιση με την $S_H(DT_p, DT_{100})$. Για τα περισσότερα από τα σύνολα δεδομένων, η συσχέτιση αυτή υπερβαίνει το 0.9, με μόνη εξαίρεση το σύνολο δεδομένων *waveform-5000* για το οποίο παρουσιάζεται χαμηλή συσχέτιση. Η $S_{P^U_{D_1 \cup D_2}}$ παρουσιάζει επίσης ισχυρή συσχέτιση με την $S_H(DT_p, DT_{100})$, αν και όχι τόσο ισχυρή όσο με την $S_{P^Q_{D_1 \cup D_2}}$. Το σύνολο δεδομένων *waveform-5000* αποτελεί και πάλι εξαίρεση καθώς εμφανίζει το μεγαλύτερο βαθμό συσχέτισης.

Οι συντελεστές συσχέτισης Pearson εκτιμούν ανα υπάρχει γραμμική συσχέτιση μεταξύ δύο μεταβλητών, ωστόσο δεν δείχνουν πόσο καλή εκτίμηση είναι η μία μεταβλητή για την άλλη. Αυτό είναι ιδιαίτερα εμφανές στην περίπτωσή μας καθώς η έννοια της γραμμικής συσχέτισης του $S_{P^Y_X}(DT_p, DT_{100})$ με το $S_H(DT_p, DT_{100})$ είναι διαφορετική για τα διάφορα σύνολα δεδομένων όπως φαίνεται στις Εικόνες 5.6, 5.7. Προκειμένου να εκτιμήσουμε την αξία πρόβλεψης των διαφόρων $S_{P^Y_X}(DT_p, DT_{100})$ όσον αφορά στο $S_H(DT_p, DT_{100})$, υπλογίζουμε τη *μέση απόλυτη απόκλιση* (Mean Absolute Deviation - MAD). Η τιμή MAD δύο μεταβλητών a και b για τις οποίες έχουμε N ζεύγη παρατηρήσεων δίνεται από τον τύπο:

$$MAD(a, b) = \sum_i^N \frac{|a_i - b_i|}{N},$$

Οι τιμές MAD για τα πειράματά μας φαίνονται στον Πίνακα 5.4. Τα αποτελέσματα δείχνουν την καλή ικανότητα πρόβλεψης της $S_{P^Q_{D_1 \cup D_2}}$ - το μέσο λάθος της στην πρόβλεψη της $S_H(DT_p, DT_{100})$ είναι 0.1. Η απόδοση της $S_{P^U_{D_1 \cup D_2}}$ είναι αξιοσημείωτα χειρότερη, καθώς η μέση τιμή MAD αυτής είναι ις περίπου 0.3.

Ο στόχος της παρούσας ενότητας ήταν να συγκρίνει και να αξιολογήσει τις διάφορες εκδόσεις του μέτρου σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης. Οι διάφορες εκδόσεις είναι αποτέλεσμα των διαφορετικών υποθέσεων

σύνολο δεδομένων	$S_{P^U_{D_1 \cup D_2}}$	$S_{P^Q_{D_1 \cup D_2}}$	$S_{P^Q_H}$
mfeat-factors	0.692	0.971	0.993
mfeat-fourier	0.852	0.927	0.999
mfeat-karhunen	0.858	0.910	0.999
mfeat-zernike	0.869	0.911	0.987
segment-challenge	0.831	0.951	0.986
waveform-5000	0.969	0.712	0.998

Πίνακας 5.3: Οι Pearson συντελεστές συσχέτισης της $S_{P^Y_X}(DT_p, DT_{100})$ με την $S_H(DT_p, DT_{100})$

σύνολο δεδομένων	$S_{P^U_{D_1 \cup D_2}}$	$S_{P^Q_{D_1 \cup D_2}}$	$S_{P^Q_H}$
μφεατ-φαστορς	0.504	0.063	0.014
μφεατ-φουριερ	0.301	0.114	0.015
μφεατ-καρηυνεν	0.279	0.158	0.013
μφεατ-ζερνικε	0.316	0.108	0.022
σεγμεντ-σηαλλενγε	0.289	0.016	0.005
ωαεφορμ-5000	0.120	0.140	0.003
Μέση τιμή	0.302	0.1	0.012

Πίνακας 5.4: Μέση απόλυτη απόκλιση μεταξύ $S_{P^Y_X}(DT_p, DT_{100})$ και $S_H(DT_p, DT_{100})$

σχετικά με την κατανομή που "κυβερνάει" το χώρο των γνωρισμάτων. Για να υπολογίσουμε τη σημασιολογική ομοιότητα δύο δέντρων απόφασης θα πρέπει να μπορούμε να υποθέσουμε μία συγκεκριμένη κατανομή $P(A)$ σχετικά με το χώρο των γνωρισμάτων. Η "επικάλυψη" των δύο δέντρων παρέχει μία πιο αναλυτική τμηματοποίηση του χώρου γνωρισμάτων και η σημαντικότητα της συμφωνίας (ή διαφωνίας) των δύο δέντρων σε μία συγκεκριμένη περιοχή της επικάλυψης εξαρτάται από την πυκνότητα της περιοχής αυτής με βάση την κατανομή $P(A)$. Αν, για παράδειγμα, τα δύο δέντρα διαφωνούν σε μία συγκεκριμένη περιοχή για την οποία η πυκνότητα με βάση την $P(A)$ είναι μηδέν, αυτή η διαφωνία δεν πρόκειται να επηρεάσει την ομοιότητα των δύο δέντρων ακόμα και αν ο όγκος της εν λόγω περιοχής είναι μεγάλος. Εναλλακτικά, αν δεν θέλουμε να υποθέσουμε κάποια συγκεκριμένη κατανομή σχετικά με το χώρο των γνωρισμάτων και θέλουμε να υπολογίσουμε την ομοιότητα σε όλους τους πιθανούς κόσμους θα πρέπει να υιοθετήσουμε την υπόθεση της ομοιόμορφης κατανομής στο χώρο των γνωρισμάτων (και αυτή η περίπτωση καλύπται από το προτεινόμενο πλαίσιο).

Προκειμένου να αξιολογήσουμε το προτεινόμενο μέτρο σημασιολογικής ομοιότητας, χρησιμοποιήσαμε την έννοια της εμπειρικής σημασιολογικής ομοιότητας, η οποία υπολογίζεται σε ένα ανεξάρτητο hold out σύνολο. Η απόδοση των διαφορετικών εκδόσεων του προτεινόμενου μέτρου σημασιολογικής ομοιότητας εξαρτάται από το πόσο διαφέρει η εκτίμηση της $P(A)$ που χρησιμοποιούν από την $P(A)$ που "κυβερνά". το hold out σύνολο, πάνω στο οποίο υπολογίζεται η εμπειρική σημασιολογική ομοιότητα. Η επιλογή της κατάλληλης $P(A)$ θα πρέπει να γίνει με βάση τη γνώση που έχουμε για το εκάστοτε πεδίο εφαρμογής. Αν γνωρίζουμε πως το

πρόβλημα κατηγοριοποίησης "κυβερνάται" από μία συγκεκριμένη $P(A)$, τότε θα πρέπει να χρησιμοποιήσουμε αυτή την $P(A)$ στο μέτρο μας. Εναλλακτικά, αν δεν έχουμε τέτοιου είδους γνώση, μπορούμε να εκτιμήσουμε την $P(A)$ από τα σύνολα δεδομένων εκπαίδευσης από τα οποία προέρχονται τα δέντρα απόφασης, όπως π.χ. στην $S_{P^Q_{D_1 \cup D_2}}$.

5.5 Σχετικές εργασίες

Αν και για τη δημιουργία δέντρων απόφασης έχει προταθεί ένας μεγάλος αριθμός αλγορίθμων, λίγες είναι οι εργασίες που ασχολούνται με το θέμα της σύγκρισης δέντρων απόφασης και ειδικότερα με το θέμα της αποτίμησης της σημασιολογικής ομοιότητας των δέντρων απόφασης, εξαίρεση αποτελεί η [78]. Τελευταία, έχουν προταθεί διάφορες προσεγγίσεις που χρησιμοποιούν μοντέλα δέντρων απόφασης για να συγκρίνουν πρωτογενή σύνολα δεδομένων, π.χ., [25, 67].

Ο Turney [78], παρουσίασε ένα πλαίσιο για την αποτίμηση της *ευστάθειας* ενός αλγορίθμου κατηγοριοποίησης, δηλαδή του βαθμού με τον οποίο ο αλγόριθμος παράγει επαναληπτικά αποτελέσματα όταν εκπαιδεύεται σε διαφορετικά σύνολα δεδομένων, τα οποία ωστόσο προέρχονται από την ίδια γεννήτρια κατανομή. Δεδομένου ότι τα σύνολα δεδομένων προέρχονται από την ίδια γεννήτρια, θα πρέπει ο αλγόριθμος να εξάγει τις ίδιες έννοιες από τα διάφορα σύνολα δεδομένων. Για να μετρήσει την ευστάθεια, ο Turney εισήγαγε ένα σημασιολογικό μέτρο της ομοιότητας και το ονόμασε *συμφωνία* (agreement). Η συμφωνία δύο ταξινομητών ορίζεται ως η πιθανότητα να προβλέψουν την ίδια κλάση για κάποιο τυχαίο στιγμιότυπο που προέρχεται από την κατανομή που "γέννησε" τα σύνολα δεδομένων των ταξινομητών. Ο Turney υπολογίζει εμπειρικά την συμφωνία δύο ταξινομητών, εφαρμόζοντας τους σε τεχνητά σύνολα δεδομένων ελέγχου που προέρχονται από την $U(A)$, την ομοιόμορφη κατανομή πάνω στο χώρο των γνωρισμάτων. Να σημειώσουμε ότι, βάσει του Turney, η συμφωνία υπολογίζεται πάνω σε στιγμιότυπα της $U(A)$ και όχι της $P(A, C)$ (που είναι η απο κοινού κατανομή των γνωρισμάτων πρόβλεψης και του γνωρίσματος κλάσης). Ο λόγος πίσω από την επιλογή της $U(A)$ είναι ότι η συμφωνία πρέπει να εξεταστεί σε όλους τους πιθανούς κόσμους εισόδου. Όπως δικαιολογείται στην [78], χρησιμοποιώντας την $U(A)$ αποκλείονται τυχόν στατιστικές σχέσεις μεταξύ των γνωρισμάτων, οι οποίες ενδέχεται να υπάρχουν στην κατανομή $P(A, C)$. Να σημειώσουμε πως η συμφωνία, όπως εξετάζεται από τον Turney [78], απαιτεί την αποτίμηση των δέντρων απόφασης με βάση στιγμιότυπα που προέρχονται από την κατανομή $U(A)$.

Τελευταία, έχουν προταθεί αρκετές μέθοδοι εντοπισμού αλλαγών που χρησιμοποιούν τη σύγκριση δέντρων απόφασης για τη *σύγκριση συνόλων από πρωτογενή δεδομένα* (dataset comparison). Η ιδέα πίσω από αυτές τις προσεγγίσεις είναι ότι τα δέντρα απόφασης αναπαριστούν ενδιαφέροντα χαρακτηριστικά των πρωτογενών δεδομένων, συνεπώς μπορούν να χρησιμοποιηθούν για τη σύγκριση τους. Όλες οι μέθοδοι στην κατηγορία αυτή ακολουθούν την ίδια λογική: χρησιμοποιούν τα δέντρα απόφασης προκειμένου να δημιουργήσουν μία πιο εκλεπτυσμένη δομή και στη συνέχεια συγκρίνουν τις κατανομές των δύο συνόλων δεδομένων πάνω σε αυτή την (κοινή) εκλεπτυσμένη δομή. Στη συνέχεια, περιγράφουμε κάποιες αντιπροσωπευτικές μεθόδους στην κατηγορία αυτή.

Στην [25], οι Ganti et al. πρότειναν το πλαίσιο FOCUS για τον υπολογισμό της διαφοράς μεταξύ δύο συνόλων δεδομένων D_1, D_2 με βάση τα μοντέλα εξόρυξης γνώσης που εξάγονται από τα δεδομένα αυτά. Τα δέντρα απόφασης είναι μεταξύ

των μοντέλων εξόρυξης γνώσης που μελετούνται στην εργασία αυτή. Έστω DT_1, DT_2 είναι τα δέντρα απόφασης που προέρχονται από τα σύνολα δεδομένων D_1, D_2, αντίστοιχα. Κάθε δέντρο απόφασης διαμερίζει το χώρο γνωρισμάτων σε ένα σύνολο από μη-επικαλυπτόμενες περιοχές, όπου κάθε περιοχή αντιστοιχεί σε ένα κόμβο-φύλλο του δέντρου. Αρχικά, μία πιο εκλεπτυσμένη δομή, καλούμενη Greatest Common Refinement - GCR), δημιουργείται από την επικάλυψη των τμηματοποιήσεων που τα δέντρα απόφασης ορίζουν πάνω στο χώρο γνωρισμάτων. Στη συνέχεια, τα στιγμιότυπα των D_1, D_2 προβάλλονται πάνω στο GCR και έτσι υπολογίζεται, για κάθε σύνολο δεδομένων, το πλήθος των στιγμιοτύπων που πέφτει σε κάθε περιοχή του GCR. Η διαδικασία αυτή απαιτεί την εκτέλεση επερωτήσεων στα αρχικά σύνολα D_1 και D_2. Στη συνέχεια, η διαφορά μεταξύ των δύο συνόλων υπολογίζεται αθροίζοντας, για κάθε περιοχή του GCR, τη διαφορά στο πλήθος των στιγμιοτύπων που καταλήγουν στην περιοχή αυτή για κάθε σύνολο δεδομένων. Όπως φαίνεται από την περιγραφή αυτή, ο $FOCUS$ χρειάζεται πρόσβαση στα πρωτογενή δεδομένα προκειμένου να υπολογίσει τη διαφορά τους.

Στην [80], οι Wang και Pei ασχολούνται με το πρόβλημα της ποσοτικοποίησης των διαφορών μεταξύ δύο συνόλων δεδομένων κατηγοριοποίησης D_1, D_2. Η ιδέα είναι να αναθέσουμε μία *υπογραφή* (signature) σε κάθε σύνολο δεδομένων και στη συνέχεια να συγκρίνουμε τα σύνολα αυτά με βάση τις υπογραφές τους. Η εύρεση μιας κάλής υπογραφής που θα μπορεί να χρησιμοποιηθεί ως (κοινή) βάση για τη σύγκριση αποτελεί πρόκληση. Η προφανής λύση είναι να χρησιμοποιήσουμε για το σκοπό αυτό την τμηματοποίηση που ένα από τα δύο δέντρα, τα οποία εξάγονται από τα D_1, D_2 σύνολα, δημιουργεί πάνω στο χώρο των γνωρισμάτων. Ωστόσο, όπως αναφέρουν οι συγγραφείς, αυτή η προσέγγιση είναι ελαττωματική καθώς η κατανομή από την οποία προέρχεται το D_1 μπορεί να είναι εντελώς διαφορετική από την κατανομή από την οποία προέρχεται το D_2. Συνεπώς, μία τέτοια προσέγγιση θα ήταν υποκειμενική ως προς το σύνολο δεδομένων από το οποίο προέρχεται το δέντρο απόφασης που θα χρησιμοποιηθεί ως κοινή βάση για τη σύγκριση. Για παράδειγμα, αν χρησιμοποιούσαμε την τμηματοποίηση του DT_1 για τη σύγκριση, το αποτέλεσμα θα ήταν υποκειμενικό ως προς το σύνολο δεδομένων D_1. Για το σκοπό αυτό, οι συγγραφείς προτείνουν να χρησιμοποιήσουμε ως υπογραφή μία *αυθαίρετη δεντρική δομή* που τμηματοποιεί τον πολυδιάστατο χώρο του προβλήματος σε ένα πλήθος από κουτιά (bin).

Αντί για μία μόνο τμηματοποίηση οι συγγραφείς προτείνουν τη χρήση πολλαπλών τμηματοποιήσεων, όπου κάθε τμηματοποίηση έχει δημιουργεί ανεξάρτητα από τις άλλες και με τυχαίο τρόπο. Οι συγγραφείς προτείνουν δύο τρόπους για τη δημιουργία τυχαίων τμηματοποιήσεων: τα τυχαία δάση (random forests) και τα τυχαία ιστογράμματα (random histograms). Από τη μελέτη τους φαίνεται ότι η δομή των τυχαίων ιστογραμμάτων αποτελεί καλύτερη λύση στο πρόβλημα της εύρεσης μιας κοινής δομής για τη σύγκριση και αυτό οφείλεται στο γεγονός ότι τα τυχαία ιστογράμματα είναι πιο ποικιλόμορφα (diverse) σε σχέση με τα τυχαία δάση. Για να μετρήσουν την ποικιλομορφία οι συγγραφείς χρησιμοποιούν το πλήθος των διαφορετικών συνδιασμών γνωρισμάτων. Η ποικιλομορφία είναι μία επιθυμητή ιδιότητα καθώς εγγυάται μεγαλύτερη εξερεύνηση του χώρου του προβλήματος και έτσι εξασφαλίζει ότι η υπογραφή θα "ταιριάζει" σε κάθε σύνολο δεδομένων. Το επόμενο βήμα, μετά την κατασκευή της (κοινής) δομής πάνω στην οποία θα λάβει χώρα η σύγκριση, είναι το "γέμισμα" της υπογραφής με στιγμιότυπα από τα δύο σύνολα δεδομένων. Στη συνέχεια, η απόσταση μεταξύ των δύο συνόλων δεδομένων υπολογίζεται συναθροίζοντας τις διαφορές τους πάνω σε κάθε ένα από τα N διαφορετικά τυχαία ιστογράμματα. Για τη σύγκριση δύο τυχαί-

ων ιστογραμμάτων, οι συγγραφείς χρησιμοποιούν την απόσταση Manhattan. Να τονίσουμε πως η μέθοδος αυτή απαιτεί πρόσβαση στα αρχικά σύνολα δεδομένων προκειμένου να υπολογίσει τη διαφορά των συνόλων αυτών.

Πρόσφατα, η Pekerskaya et al. [67] προτείνουν μία μέθοδο για την εξόρυξη *μεταβαλλόμενων περιοχών* (changing regions) από σύνολα δεδομένων στα οποία δεν επιτρέπεται η πρόσβαση. Μία περιοχή χαρακτηρίζεται ως μεταβαλλόμενη αν εμφανίζεται με διαφορετικές ετικέτες κλάσεων στα δύο σύνολα δεδομένων. Ο στόχος είναι η εύρεση τέτοιων περιοχών και η ταξινόμησή τους με βάση το ρυθμό αλλαγής, χωρίς ωστόσο να προσπευλαύνονται τα αρχικά σύνολα δεδομένων. Ένας τέτοιος περιορισμός μπορεί να είναι απαραίτητος για λόγους ιδιωτικότητας ή εξαιτίας της έλλειψης πρόσβασης στα πρωτογενή δεδομένα (για παράδειγμα, στα ρεύματα δεδομένων τα παλιά δεδομένα ξεχνιούνται μετά από μερικές χρονικές στιγμές). Οι συγγραφείς δικαιολογούν ότι η τμηματοποίηση που ένα δέντρο απόφασης δημιουργεί πάνω στο χώρο γνωρισμάτων δεν αποτελεί καλή προσέγγιση της κατανομής του συνόλου δεδομένων. Για το σκοπό αυτό, επεκτείνουν το κλασσικό μοντέλο δέντρων απόφασης διασπώντας περαιτέρω κάθε κόμβο-φύλλο του δέντρου απόφασης σε ένα σύνολο από συστάδες, μέσω κάποιου αλγορίθμου συσταδοποίησης. Το νέο μοντέλο, που ονομάζεται δέντρο απόφασης ενσωματωμένο με συστάδες (cluster-embedded decision tree, παρέχει μία καλύτερη προσέγγιση της κατανομής του συνόλου δεδομένων σε σχέση με την προσέγγιση που παρέχει ένα (απλό) δέντρο απόφασης. Μετά την εξαγωγή των cluster-embedded δέντρων απόφασης, υπολογίζεται η επικάλυψη των δύο αυτών δομών και εν συνεχεία, ακολουθώντας μία λογική παρόμοια με αυτή του FOCUS [25], υπολογίζουμε τα στατιστικά των δύο συνόλων δεδομένων πάνω στην κοινή αυτή δομή. Για τον υπολογισμό των στατιστικών στηρίζονται μόνο στα στοιχεία των αντίστοιχων cluster-embedded δέντρων απόφασης. Η υπόθεση που κάνουν είναι ότι τα στιγμιότυπα είναι ομοιόμορφα κατανεμημένα μέσα στις συστάδες του cluster-embedded δέντρου απόφασης. Να επισημάνουμε πάλι πως αυτή η μέθοδος απαιτεί τη δημιουργία των cluster-embedded δέντρων απόφασης και δεν χρησιμοποιεί τα ίδια τα μοντέλα δέντρων απόφασης.

Υπάρχει μεγάλο πλήθος εργασιών στη σύγκριση της δομής δύο δέντρων με βάση την *απόσταση επεξεργασίας* (edit distance), π.χ., [86]. Οι προσεγγίσεις αυτές βασίζονται στην μέτρηση του πλήθους και του κόστους των λειτουργιών επεξεργασίας (εισαγωγή, διαγραφή, ενημέρωση) που απαιτούνται προκειμένου να μετατρέψουμε ένα δέντρο σε ένα άλλο. Ωστόσο, οι μέθοδοι αυτές δουλεύουν με συμβολικά δέντρα όπου οι κόμβοι μαρκάρονται με σύμβολα από ένα δοθέν αλφάβητο. Στην περίπτωση των δέντρων απόφασης όμως, οι κόμβοι είναι πιο πολύπλοκοι καθώς περιλαμβάνουν συνθήκες πάνω στα σύμβολα-γνωρίσματα και επιπλέον, σε κάθε μονοπάτι του δέντρου απόφασης ανατίθεται ένα βάρος, το οποίο καθορίζεται από το πλήθος των στιγμιοτύπων που ακολουθούν αυτό το μονοπάτι.

Στην προσέγγισή μας αξιοποιούμε την πληροφορία που παρέχει ένα μοντέλο δέντρου απόφασης προκειμένου να ορίσουμε μέτρα απόστασης μεταξύ δέντρων απόφασης και συνόλων δεδομένων κατηγοριοποίησης. Ανάλογα με το αν επιτρέπεται η πρόσβαση στα πρωτογενή δεδομένα ή όχι προτείνουμε διαφορετικούς τρόπους για την αποτίμηση της ομοιότητας.

5.6 Συμπεράσματα

Στο κεφάλαιο αυτό, παρουσιάσαμε ένα γενικό πλαίσιο ομοιότητας το οποίο χρησιμοποιεί τα δέντρα απόφασης για την αποτίμηση της ομοίότητας μεταξύ δέντρων απόφασης και συνόλων δεδομένων κατηγοριοποίησης. Η *ομοιότητα δύο δέντρων απόφασης* υπολογίζεται με βάση τη συμφωνία που παρουσιάζουν σχετικά με τις κλάσεις που προβλέπουν για στιγμιότυπα που προέρχονται από το χώρο γνωρισμάτων του προβλήματος, και αντιστοιχεί στην ένοια της εμπειρικής σημασιολογικής ομοιότητας (empirical semantic similarity) [78]. Ο υπολογισμός της *ομοιτητας δύο συνόλων δεδομένων* μπορεί να γίνει με βάση την κατανομή τους στο χώρο των γνωρισμάτων $P(A)$, με βάση την απο κοινού κατανομή γνωρισμάτων και κλάσης $P(A, C)$ ή με βάση την υπό-συνθήκη κατανομή των κλάσεων με βάση τα γνωρίσματα $P(C|A)$. Όλα τα παραπάνω αποτελούν ειδικές περιπτώσεις του προτεινόμενου πλαισίου.

Στη δουλειά αυτή, εστιάζουμε στην αποτίμηση της σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης, δηλαδή, υπολογίζουμε το βαθμό στον οποίο τα δέντρα απόφασης συμφωνούν όσον αφορά στις προβλέψεις τους πάνω στο χώρο των γνωρισμάτων. Κρίσιμο σημείο σε αυτή τη διαδικασία αποτελεί η επιλογή μιας κατάλληλης κατανομής $P(A)$ για το χώρο των γνωρισμάτων, με βάση την οποία θα γίνει η σύγκριση. Η επιλογή της κατανομής $P(A)$ αντανακλά το τι πιστεύουμε σχετικά με τον πραγματικό κόσμο όπου θα εφαρμοστούν τα δέντρα απόφασης. Αν δεν έχουμε κάποια πρότερη γνώση σε σχέση με αυτή την κατανομή, θα μπορούσαμε να χρησιμοποιήσουμε την ομοιόμορφη κατανομή $U(A)$ για την $P(A)$, εξετάζοντας έτσι τη σημασιολογική ομοιότητα δύο δέντρων απόφασης σε όλους τους πιθανούς κόσμους εισόδου.

Πειραματιστήκαμε με διάφορους τρόπους εκτίμησης της της κατανομής του χώρου των γνωρισμάτων $P(A)$ και συγκρίναμε τα διάφορα στιγμιότυπα του μέτρου σημασιολογικής ομοιότητας δέντρων απόφασης που προτείνουμε με το μέτρο της εμπειρικής σημασιολογικής ομοιότητας, το οποίο προκύπτει εφαρμόζοντας τα δέντρα απόφασης σε ανεξάρτητα hold out σύνολα ελέγχου. Το πόσο καλά η υπολογιζόμενη σημασιολογική ομοιότητα συσχετίζεται με την (πραγματική) εμπειρική σημασιολογική ομοιότητα εξαρτάται από την γνώση που έχουμε σχετικά με την κατανομή $P(A)$ από την οποία προέρχεται το ανεξάρτητο hold out σύνολο ελέγχου. Πιο συγκεκριμένα, όταν υπολογίζουμε την $P(A)$ μέσω επερωτήσεων στα πραγματικά σύνολα δεδομένων, τότε η υπολογιζόμενη σημασιολογική ομοιότητα $S_{P^Q_{D_1 \cup D_2}}$ συσχετίζεται πάρα πολύ καλά με την πραγματική σημασιολογική ομοιότητα. Μάλιστα, περιμένουμε η τιμή της $S_{P^Q_{D_1 \cup D_2}}$ να συγκλίνει στην πραγματική τιμή της σημασιολογικής ομοιότητας όσο το μέγεθος των συνόλων δεδομένων μεγαλώνει, καθώς η εκτίμηση της $\mathbf{P(A)}$ θα συγκλίνει στην πραγματική $P(A)$.

Πιστεύουμε πως η μεγαλύτερη αξία ενός μέτρου σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης έγκειται στη δυνατότητα που προσφέρει να καθορίσουμε κατά πόσο οι παρατηρούμενες διαφορές μεταξύ δύο δέντρων απόφασης είναι απλά επιφανειακές διαφορές στη δομής τους ή ανταποκρίνονται σε πραγματικές σημασιολογικές διαφορές των εννοιών που περιγράφονται μέσω των δέντρων αυτών και βεβαιώς να ποσοτικοποιήσουμε αυτές τις διαφορές. Αυτό είναι ένα πρόβλημα που "ταλαιπωρεί" τα δέντρα απόφασης εξαιτίας της μεγάλης τους ευαισθησίας σε έστω και μικρές αλλαγές στο σύνολο δεδομένων εκπαίδευσης από το οποίο έχουν εξαχθεί.

Πρώιμες εκδόσεις αυτής της μελέτης υπάρχουν στις [55, 56].

5.7 Ανοιχτά θέματα

Η ύπαρξη ενός μέτρου ομοιότητας μεταξύ μοντέλων κατηγοριοποίηση, δέντρων απόφασης στην περίπτωσή μας, μας επιτρέπει να εκτελέσουμε διάφορες εργασίες εξόρυξης γνώσης πάνω στα μοντέλα αυτά αντί για τα πρωτογενή δεδομένα. Αυτό είναι ένα είδος *μετα-εξόρυξης ή μετα-ανάλυσης*. Στη συνέχεια παρουσιάζουμε διάφορες άλλες ερευνητικές κατευθύνσεις.

Συσταδοποίηση δέντρων απόφασης Χρησιμοποιώντας το μέτρο σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης, θα μπορούσαμε να ομαδοποιήσουμε ένα σύνολο δέντρων απόφασης σε ένα πλήθος συστάδων και να βρούμε το *αντιπροσωπευτικό δέντρο απόφασης* (representative decision tree) για κάθε συστάδα. Τα δέντρα απόφασης που ανήκουν στην ίδια συστάδα θα πρέπει να έχουν σημασιολογική ομοιότητα, δηλαδή (σε κάποιο βαθμό) θα πρέπει να συμφωνούν στις προβλέψεις τους σχετικά με το πρόβλημα κατηγοριοποίησης που περιγράφουν. Μία τυπική εφαρμογή είναι σε περιπτώσεις *κατενεμημένης εξόρυξης γνώσης* (βλέπε το παράδειγμα της τράπεζας στην Ενότητα 5.1).

Όπως και στην κλασική περίπτωση συσταδοποίησης πάνω σε πρωτογενή δεδομένα, η συσταδοποίηση δέντρων απόφασης περιλαμβάνει τα ακόλουθα βήματα: ι) τον ορισμό ενός μέτρου ομοιότητας μεταξύ δέντρων απόφασης, ιι) την επιλογή ενός αλγορίθμου συσταδοποίησης και ιιι) τον ορισμό ενός κριτηρίου ποιότητας για την αξιολόγηση της συσταδοποίησης που προκύπτει. Για το βήμα ι), θα μπορούσαμε να χρησιμοποιήσουμε το μέτρο σημασιολογικής ομοιότητας που προτείνουμε σε αυτό το κεφάλαιο.

Παρακολούθηση της εξέλιξης των δέντρων απόφασης σε δυναμικά περιβάλλοντα Δεδομένου ότι στις μέρες μας τα δεδομένα είναι κυρίως δυναμικά (π.χ., ρεύματα δεδομένων, δεδομένα από σένσορες) υπάρχει η ανάγκη για την παρακολούθηση της εξέλιξης τους. Ένας τρόπος να γίνει αυτό είναι μελετώντας πως εξελίσονται τα μοντέλα εξόρυξης γνώσης που εξάγονται από τα δεδομένα αυτά.

Έτσι θα μπορούσαμε να μελετήσουμε τις αλλαγές σε ένα δυναμικό πληθυσμό μελετώντας πως εξελίσονται τα δέντρα απόφασης που εξάγονται από τον πληθυσμό αυτό στην πορεία του χρόνου. Διάφορες μεταβολές θα μπορούσαν να εντοπιστούν: Εξακολουθεί το παλιό δέντρο να περιγράφει καλά τα νέα δεδομένα ή μήπως υπάρχει κάποια άλλη τάση στα νέα δεδομένα; Υπάρχουν περιοχές απόφασης που δεν μεταβάλλονται σε σχέση με την κλάση που προβλέπουν; Υπάρχουν περιοχές που αλλάζουν συνεχώς τις προβλέψεις τους;

Η παρακολούθηση βασίζεται σε κάποια συνάρτηση ομοιότητας που αξιολογεί πόσο όμοια είναι δύο διαδοχικά μοντέλα δέντρων απόφασης. Ωστόσο, υπάρχουν και άλλα θέματα που πρέπει να διερευνηθούν όπως ο αποδοτικός υπολογισμός της ομοιότητας, η αποδοτική παρουσίαση της εξέλιξης στον τελικό χρήστη κ.ο.κ.

Απλοποίηση συνόλων από δέντρα απόφασης Ένα *σύνολο/ δάσος από δέντρα απόφασης* (decision tree ensemble/ forest) είναι μία συλλογή από δέντρα απόφασης που αναφέρονται στο ίδιο πρόβλημα κατηγοριοποίησης. Σε μία τέτοια δομή, τα επιμέρους δέντρα απόφασης αναπτύσσονται ανεξάρτητα. Όταν ένα νέο στιγμιότυπο εμφανίζεται, οι προβλέψεις των επιμέρους δέντρων συνδιάζονται προκειμένου να παράξουν την τελική πρόβλεψη.

Χρησιμοποιώντας ένα σύνολο από δέντρα απόφασης, αντί ενός απλού δέντρου απόφασης, η πρόβλεψη που προκύπτει είναι πιο ακριβής καθώς λόγω των πολλαπλών δέντρων απόφασης υπάρχουν καλύτερες πιθανότητες εξερεύνησης όλου του χώρου γνωρισμάτων. Συνεπώς, η ικανότητα γενίκευσης μιας τέτοιας δομής σε σχέση με ένα απλό δέντρο απόφασης είναι υψηλότερη. Ωστόσο, ένα σύνολο από δέντρα απόφασης δημιουργεί μία πιο πολύπλοκη δομή σε σύγκριση με ένα απλό δέντρο απόφασης. Συνεπώς, είναι πιο δύσκολο για το χρήστη να κατανοήσει ένα τέτοιο σύνολο, καθώς θα πρέπει να κατανοήσει όλα τα επιμέρους δέντρα που το συνθέτουν. Η ιδέα είναι να απλοποιήσουμε τη δομή αυτή μειώνοντας το πλήθος των δέντρων απόφασης που χρειάζεται να μελετήσει ο χρήστης μέσω της συσταδοποίησης, δηλαδή να οργανώσουμε τα δέντρα απόφασης σε ομάδες με παρόμοια δέντρα.

Το προφανές πλεονέκτημα του να έχουμε ένα μόνο ή λίγα δέντρα απόφασης αντί ενός συνόλου από δέντρα απόφασης είναι η πολύ πιο εύκολη ερμηνεία αυτού του δέντρου από τον τελικό χρήστη.

Κεφάλαιο 6

Σύγκριση Συστάδων (και Συσταδοποιήσεων) - Εντοπισμός και Διαχείριση Μεταβολών σε Δυναμικά Περιβάλλοντα

Στο κεφάλαιο αυτό, χρησιμοποιούμε την έννοια της ομοιότητας μεταξύ συστάδων και συσταδοποιήσεων προκειμένου να παρακολουθήσουμε την εξέλιξή τους σε ένα δυναμικό περιβάλλον και να εντοπίσουμε τυχόν μεταβολές μεταξύ των συστάδων.

Το τρέχον κεφάλαιο έχει οργανωθεί ως εξής: Το πρόβλημα και τα κίνητρα της μελέτης περιγράφονται στην Ενότητα 6.1. Το μοντέλο του δυναμικού περιβάλλοντος περιγράφεται στην Ενότητα 6.2. Στην Ενότητα 6.3, παρουσιάζουμε το πλαίσιο $MONIC$ για τη μοντελοποίηση και τον εντοπισμό των αλλαγών/μεταβολών σε συστάδες που ορίζονται ανεξάρτητα από τον τύπο συστάδας στον οποίο ανήκουν (δηλαδή, ιεραρχικός, διαμεριστικός, βάσει πυκνότητας). Επεκτείνουμε το πλαίσιο $MONIC$ στο $MONIC+$ (Ενότητα 6.4), το οποίο λαμβάνει επίσης υπόψη και τα ειδικά χαρακτηριστικά κάθε τύπου συστάδας. Οι μεταβολές των συστάδων οργανώνονται σε ένα Γράφο Εξέλιξης (*Evolution Graph*) (Ενότητα 6.5), όπου οι κόμβοι αναπαριστούν συστάδες που παρατηρούνται σε διαφορετικές χρονικές στιγμές, ενώ οι ακμές αναπαριστούν μεταβολές μεταξύ των συστάδων. Στην Ενότητα 6.6, η εξέλιξη των συστάδων συνοψίζεται μέσω του πλαισίου $FINGERPRINT$ που απομακρύνει τις λιγότερο πληροφοριακές συστάδες. Η πειραματική μελέτη στα $MONIC$, $MONIC+$ και $MONIC$ παρουσιάζεται στην Ενότητα 6.7. Οι σχετικές εργασίες παρουσιάζονται στην Ενότητα 6.8. Ολοκληρώνουμε τη μελέτη μας στην Ενότητα 6.9 παρουσιάζοντας τα συμπεράσματα της μελέτης. Στην Ενότητα 6.10 περιγράφουμε διάφορα ανοιχτά ερευνητικά θέματα.

Λέξεις κλειδιά μεταβολές συστάδων, εξέλιξη συστάδων, εντοπισμός αλλαγών, σύνοψη συστάδων, δυναμικά περιβάλλοντα.

6.1 Εισαγωγή

Τα ρεύματα δεδομένων (data streams) εμφανίζονται σε πολλές σύγχρονες εφαρμογές και επιβάλλουν νέες προκλήσεις σχετικά με τη διαχείρισή τους λόγω του μεγέθους τους και της δυναμικότητάς τους. Μία από αυτές τις προκλήσεις είναι ο αποδοτικός εντοπισμός και η παρακολούθηση των αλλαγών στον πληθυσμό. Γενικότερα, η παρακολούθηση των αλλαγών είναι βασική για εφαρμογές που απαιτούν μακροχρόνια πρόβλεψη και δράση.

Τα μοντέλα συστάδων χρησιμοποιούνται συχνά για τη μελέτη της δυναμικότητας ενός πληθυσμού. Τα τελευταία χρόνια μάλιστα, εξαιτίας της δυναμικής φύσης των δεδομένων, έχει αναγνωριστεί ότι οι συστάδες ενός συνόλου δεδομένων επηρεάζονται από τις αλλαγές στον πληθυσμό από τον οποίο εξάγονται οι συστάδες. Ένα μεγάλο πλήθος εργασιών αφορά στην *προσαρμογή* των συστάδων στον μεταβαλλόμενο πληθυσμό. Πρόσφατα ωστόσο έχουν προταθεί εργασίες που σχετίζονται με το πρόβλημα του *εντοπισμού και της κατανόησης* των αλλαγών ως μέσο για να γνωρίσουμε καλύτερα τα δεδομένα μας και να υποστηρίξουμε στρατηγικές αποφάσεις.

Για την κατηγοριοποίηση και τον εντοπισμό των αλλαγών των συστάδων προτείνουμε το πλαίσιο $MONIC$. Το $MONIC$ παίρνει σαν είσοδο ένα συσσωρευμένο σύνολο δεδομένων τα στιγμιότυπα του οποίου γερνούν όπως συμβαίνει στις εφαρμογές ρευμάτων δεδομένων. Τα στιγμιότυπα συσταδοποιούνται σε διαδοχικές χρονικές στιγμές και η εξέλιξή τους παρακολουθείται. Για το σκοπό αυτό, προτείνουμε πρώτα ένα σύνολο από *μεταβολές συστάδων* (cluster transitions), όπως η επιβίωση (survival), η διάσπαση (split) και η απορρόφηση (absorption), και στη συνέχεια προτείνουμε *δείκτες μεταβολών* (transition indicators) που αποτελούν μέρος ενός αλγορίθμου για τον εντοπισμό των μεταβολών. Το $MONIC$ διαφοροποιείται από υπάρχουσες προσεγγίσεις για εντοπισμό αλλαγών (π.χ., [1, 26, 46]), καθώς δεν εξαρτάται από τον αλγόριθμο συσταδοποίησης, αλλά από τα μέλη των συστάδων.

Προκειμένου να αξιοποιήσουμε τα ειδικά χαρακτηριστικά κάθε τύπου συστάδας, επεκτείνουμε το πλαίσιο $MONIC$ στο πλαίσιο $MONIC+$που λαμβάνει επίσης υπόψιν και τον τύπο των συστάδων (ιεραρχικός, διαμεριστικός, με βάση την πυκνότητα) και συνεπώς, εξαρτάται από τον τύπο των προτύπων.

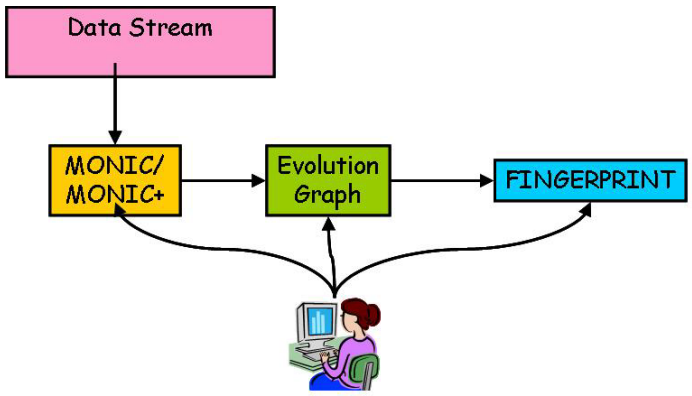
Μοντελοποιούμε τις μεταβαλλόμενες συστάδες σε μία δομή γράφου, τον *Evolution Graph*, ο οποίος αποτελείται από ακολουθίες αλλαγών από την πρώτη χρονική στιγμή που εμφανίζεται μία συστάδα μέχρι τη στιγμή που διαλύεται. Αυτός ο γράφος περιέχει πλούτο πληροφορίας σχετικά με τις μεταβολές του πληθυσμού. Υπάρχουν πολλές δυνατότητες αξιοποίησης του γράφου όπως οι επερωτήσεις και η μελέτη της ευστάθειας του πληθυσμού, μελετώντας το χρόνο ζωής των συστάδων και των συσταδοποιήσεων.

Ωστόσο, επειδή η αλλαγή και η εξέλιξη είναι μόνιμα χαρακτηριστικά των ρευμάτων δεδομένων, καθώς η περίοδος της παρατήρησης αυξάνεται ο γράφος διογκώνεται και γίνεται δύσχρηστος από την πλευρά του χρήστη. Για το λόγο αυτό προτείνουμε το πλαίσιο $FINGERPRINT$, που συμπιέζει το γράφο ώστε παρόμοιες συστάδες να συνοψίζονται με βάση κριτήρια ακρίβειας και συμπίεσης.

Συνοψίζοντας, στο κεφάλαιο αυτό μελετάμε το πρόβλημα της εξέλιξης ενός πληθυσμού σε ένα δυναμικό περιβάλλον με βάση τα μοντέλα των συστάδων που εξάγονται από τον πληθυσμό αυτό. Η προσέγγισή μας αποτελείται από τρία επιμέρους κομμάτια:

- Πρώτα, το πλαίσιο $MONIC$ εντοπίζει μεταβολές μεταξύ συστάδων που εμφανίζονται σε διαδοχικές χρονικές στιγμές.
- Στη συνέχεια, τόσο οι συστάδες όσο και οι μεταβολές αυτών οργανώνονται σε μία δομή γράφου, που καλείται *Evolution Graph*), που περιέχει όλη την ιστορία της εξέλιξης του πληθυσμού.
- Τέλος, το πλαίσιο $FINGERPRINT$ συμπιέζει τις μεταβολές αυτές σε κάποια πιο συμπαγή δομή η οποία ωστόσο παραμένει πληροφοριακή.

Αυτά τα τρία συστατικά αποτελούν την απαραίτητη δομή για τη μελέτη της εξέλιξης ενός δυναμικού πληθυσμού στην πορεία του χρόνου. Η αρχιτεκτονική της προσέγγισής μας παρατίθεται στην Εικόνα 6.1.



Σχήμα 6.1: Η αρχιτεκτονική του συστήματος

6.2 Παρακολούθηση της εξέλιξης σε δυναμικά περιβάλλοντα

Υποθέτουμε ότι το ρεύμα δεδομένων αποτελείται από ένα σύνολο εγγραφών $d_1, \ldots, d_n$ που φτάνουν τις χρονικές στιγμές $t_1, \ldots, t_n$, όπου d_i $(i = 1 \ldots n)$ είναι το υπορεύμα των δεδομένων που φτάνει στο χρονικό διάστημα $(t_{i-1}, t_i]$. Τα παλιά δεδομένα γερνούν και ξεχνιούνται βάσει μιας συνάρτησης γήρανσης των δεδομένων (data ageing function) που αναθέτει χαμηλότερα βάση σε μερικά ή όλα από τα δεδομένα του παρελθόντος, όπως συμβαίνει συνήθως στον εντοπισμό και την παρακολούθηση θεμάτων (topic detection and tracking) [6].

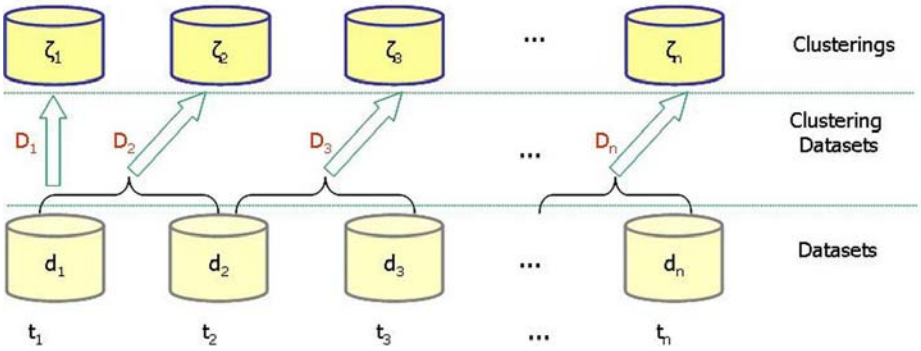
Ορισμός 5 (Συνάρτηση γήρανσης των δεδομένων) *Έστω $t_1 \ldots, t_n$ είναι οι χρονικές στιγμές της παρακολούθησης και έστω $d_i, i = 2 \ldots, n$ είναι το σύνολο των δεδομένων που συσσωρεύονται μεταξύ της t_{i-1} και t_i, όπου d_1 είναι το αρχικό σύνολο δεδομένων, ώστε $d_i \cap d_j = \emptyset$ για $i \neq j$. Μία συνάρτηση γήρανσης δεδομένων αναθέτει ένα βάρος $age(x, t_i) \in [0, 1]$ στην εγγραφή x την t_i, για κάθε $x \in \cup_{l=1}^{i} d_l$ και για κάθε t_i.*

$$age : \cup_{l=1}^{i} d_l \times \{t_1 \ldots, t_n\} \rightarrow [0, 1] \tag{6.1}$$

Η συνάρτηση αυτή καλύπτει συρόμενο παράθυρα (sliding windows) (τα βάρη των εγγραφών έξω από το παράθυρο είναι 0), αλλά και πιο περίπλοκα σχήματα, π.χ., [53] που θεωρεί επανεμφανίσεις κάθε εγγραφής και αναθέτει μεγαλύτερο βάρος σε επανεμφανιζόμενες εγγραφές.

Έστω $D_1, D_2, \ldots, D_n$ είναι τα πραγματικά σύνολα δεδομένων τις χρονικές στιγμές $t_1, \ldots, t_n$· σημειώστε πως, ανάλογα με τη συνάρτηση γήρανσης, το D_i μπορεί να περιέχει εκτός από τις εγγραφές του d_i και εγγραφές από το παρελθόν. Τα βάρη που ανατίθενται από τη συνάρτηση γήρανσης καθορίζουν την επίδραση κάθε εγγραφής στη συσταδοποίησης $\zeta_i \equiv \zeta_i(\cup_{l=1}^{i} d_l, age, t_i)$.

Ενδιαφερόμαστε να μελετήσουμε την εξέλιξη του πληθυσμού κατά μήκος του άξονα του χρόνου. Για το σκοπό αυτό, βασιζόμαστε στα μοντέλα συσταδοποίησης $\zeta_1, \zeta_2, \ldots, \zeta_n$ που εξάγονται από τα $D_1, D_2, \ldots, D_n$, αντίστοιχα. Στην Εικόνα 6.2, απεικονίζεται το μοντέλο του δυναμικού περιβάλλοντος που θεωρούμε.



Σχήμα 6.2: Παρακολούθηση δυναμικού περιβάλλοντος (μέγεθος παραθύρου = 2)

Το σύνολο των γνωρισμάτων με βάση τα οποία γίνεται η συσταδοποίηση είναι δυναμικό, συνεπώς κατά τη διάρκεια της παρατήρησης μπορεί να αλλάξει, επιτρέποντας την ενσωμάτωση κάποιου νέου γνωρίσματος και την απομάκρυνση απαρχαιωμένων γνωρισμάτων. Προκειμένου να χειριστούμε και τέτοιους *δυναμικούς χώρους γνωρισμάτων*, κάνουμε συσταδοποίηση σε κάθε χρονική στιγμή. Με τον τρόπο αυτό, οι αλλαγές μπορούν να εντοπιστούν ακόμα και αν ο θεωρούμενος χώρος γνωρισμάτων αλλάξει, δηλαδή, όταν η προσαρμογή των συστάδων δεν είναι εφικτή. Επιπλέον, αυτή η προσέγγιση επιτρέπει τον εντοπισμό αλλαγών τόσο σε ήδη εντοπισμένες συστάδες όσο και σε νέες συστάδεσ᾽

Η συσταδοποίηση πάνω σε ένα σύνολο δεδομένων μπορεί να θεωρηθεί ως η τμηματοποίηση του συνόλου δεδομένων σε ομοιογενείς ομάδες. Επικεντρωνόμαστε στην *αυστηρή συσταδοποίηση* (hard clustering), στην οποία ένα αντικείμενο ανήκει σε ακριβώς μία συστάδα, σε αντίθεση με την χαλαρή συσταδοποίηση (soft clustering) στην οποία ένα αντικείμενο ανήκει (με κάποιο ποσοστό συμμετοχής) σε παραπάνω από μία συστάδες.

Στόχος μας είναι να παρακολουθήσουμε μία συστάδα που εντοπίζεται κάποια t_i μεταξύ των συστάδων της επόμενης χρονικής στιγμής t_j. Καθώς αυτό εξαρτάται από την έννοια της ίδιας της συστάδας, παρουσιάζουμε πρώτα μία κατηγοριοποίηση των συστάδων.

Οι αλγόριθμοι συσταδοποίησης χρησιμοποιούν διάφορους ορισμούς για την έννοια των προτύπων [30]. Προτείνουμε την ακόλουθη κατηγοριοποίηση που

διευκολύνει τη μελέτη των συστάδων ως μεταβαλλόμενα αντικείμενα:

Ορισμός 6 (Συστάδες τύπου A:) *Οι συστάδες ανακαλύπτονται πάνω σε ένα μετρικό χώρο. Μία συστάδα είναι ένα γεωμετρικό αντικείμενο, π.χ., μία σφαίρα όπως στον K-means. Οι αλλαγές των συστάδων αντιμετωπίζονται ως γεωμετρικοί μετασχηματισμού.*

Ορισμός 7 (Συστάδες τύπου B1:) *Δεν υπάρχει μετρικός χώρος ή εξαρτάται από τα δεδομένα κάθε στιγμής. Μία συστάδα ορίζεται αναλυτικά ως ένα σύνολο από εγγραφές. Στην κατηγορία αυτή ανήκουν οι ιεραρχικοί αλγόριθμοι οι οποίοι χτίζουν δενδρογράμματα και εκφράζουν τις συστάδες ως σύνολα κοντινών εγγραφών. Αυτοί οι αλγόριθμοι χρησιμοποιούν ένα μετρικό χώρο για να φτιάξουν μία συσταδοποίηση σε ένα σύνολο δεδομένων, αλλά αυτός ο χώρος εξαρτάται από τα δεδομένα, με την έννοια ότι η προσθήκη μιας νέας εγγραφής μπορεί να αλλάξει τα όρια της συστάδας ακόμα και αν η εγγραφή ακόμα και αν η εγγραφή αυτή δεν ανήκει στη συστάδα.*

Ορισμός 8 (Συστάδες τύπου B2:) *Μία συστάδα περιγράφεται* διαισθητικά *σαν μία κατανομή. Για μία συστάδα X τύπου B2, συμβολίζουμε το μέγεθος του πληθυσμού του με $card(X)$, το μέσο όρο του με $\mu(X)$ και την τυπική του απόκλιση με $\sigma(X)$. Ο αλγόριθμος Expectation-Maximization - EM ανήκει σε αυτή την κατηγορία.*

Διάφοροι συνδυασμοί των ανωτέρω τύπων είναι δυνατοί, π.χ., όταν χρησιμοποιούνται τόσο τα δεδομένα όσο και στατιστικά πάνω στα δεδομένα(τύπος B1+B2). Να σημειώσουμε επίσης, πως κάθε συστάδα μπορεί να περιγραφεί αναλυτικά με βάση τα δεδομένα από τα οποία αποτελείται (δηλαδή, τύπος B1) και αυτός είναι ένας γενικός ορισμός ανεξάρτητος από τον τύπο συστάδας.

Ο Πίνακας 6.1 συνοψίζει τα σύμβολα που χρησιμοποιούνται σε αυτό το κεφάλαιο

6.3 Το πλαίσιο *MONIC* για τον εντοπισμό των μεταβολών των συστάδων

Στην Εικόνα 6.3, παρουσιάζουμε με ένα οπτικό παράδειγμα την πρόκληση της κατανόησης της αλλαγής των συστάδων σε ένα δυναμικό περιβάλλον: Παρουσιάζονται συστάδες σε δύο χρονικές στιγμές (με + και κόκκινο χρώμα συμβολίζονται οι εγγραφές που προστίθενται σε κάθε χρονική στιγμή). Οι παλιές εγγραφές ξεχνιούνται με παράμετρο παραθύρου 2. Οι συστάδες σε κάθε χρονική στιγμή φαίνονται ξεκάθαρα σε αυτό το σχήμα. Είναι επίσης προφανές ότι έχουν συμβεί αλλαγές. Είναι πολύ ενδιαφέρον να βρούμε την ίδια συστάδα ξανά και να δούμε τις αλλαγές που έχει υποστεί: "Εξαφανίστηκε κάποια συστάδα· Ή απορροφήθηκε από άλλες συστάδες· Πότε μία συστάδα παραμένει ίδια και πότε μεταλλάσσεται·" Για τον εντοπισμό τέτοιων μεταβολών, προτείνουμε το πλαίσιο *MONIC* το οποίο περιλαμβάνει ι) την κατηγοριοποίηση και ιι) των εντοπισμό των αλλαγών των συστάδων.

Το *MONIC* παίρνει σαν είσοδο δύο διαδοχικές συσταδοποιήσεις από τον (εξελισσόμενο) πληθυσμό και εξάγει τις μεταβολές μεταξύ των επιμέρους συστάδων τους. Το πρώτο βήμα για τον εντοπισμό αυτών των αλλαγών είναι να βρούμε

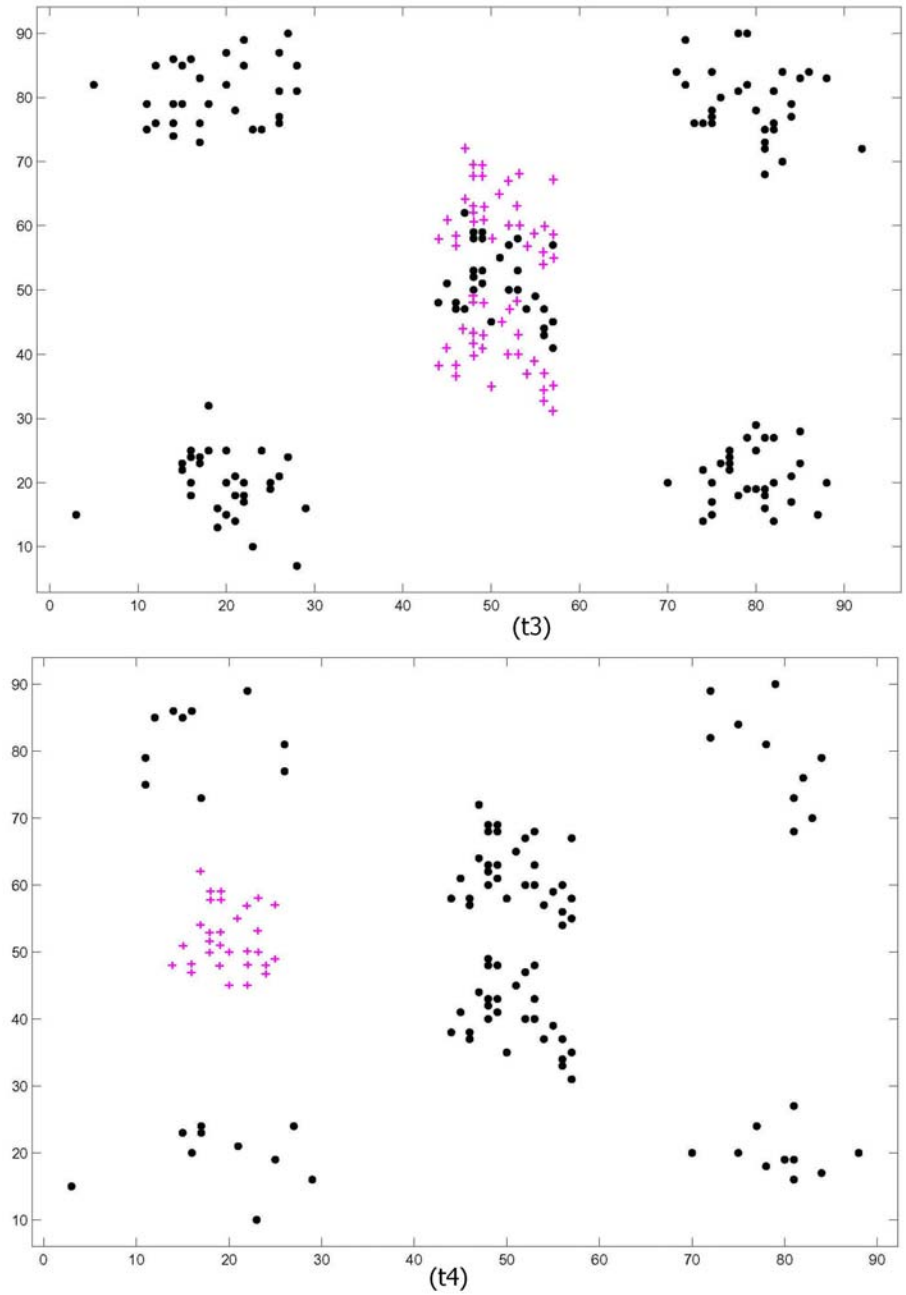
Σύμβολο	*Περιγραφή*
$age()$	η συνάρτηση γήρανσης (Ορισμός 5)
d_i	το σύνολο δεδομένων στο διάστημα $(t_{i-1}, t_i]$
D_i	το σύνολο δεδομένων την t_i με βάση και την $age()$
ζ_i	η συσταδοποίηση την t_i (βάσει του D_i)
$\|A\|$	η πληθυκότητα του συνόλου A
$overlap(X, Y)$	η επικάλυψη της X στην Y (Ορισμός 9)
$\tau \equiv \tau_{match}$	το κατώφλι ταιριάσματος $(0.5 < \tau \leq 1)$ (Ορισμός 10)
τ_{split}	το κατώφλι διάσπασης ($\tau_{split} < \tau$) (Ενότητα 6.3.2.1)
$\widehat{c}$	η ετικέτα της συστάδας c (Ενότητα 6.5.1.1)
$T \equiv trace(c)$	το μονοπάτι της αναδυόμενης συστάδας c (Ορισμός 17)
$lifetime(c)$	ο χρόνος ζωής της συστάδας c (Ορισμός 18)
$lifetime(\zeta)$	ο χρόνος ζωής της συσταδοποίησης ζ (Ορισμός 19)
$survivalRatio(\zeta)$	ο ρυθμός επιβίωσης της συσταδοποίησης ζ (Εξίσωση 6.9)
$absorptionRatio(\zeta)$	ο ρυθμός απορρόφησης της συσταδοποίησης ζ (Εξίσωση 6.9)
$passforwardRatio(\zeta)$	ο ρυθμός $passforward$ της συσταδοποίησης ζ (Εξίσωση 6.9)
$\widehat{X}$	το εικονικό κέντρο του υπο-μονοπατιού X (Ορισμός 23)
$ILoss\_trace(T, S)$	απώλεια πληροφορίας λόγω της αντικατάστασης του μονοπατιού T από τη σύνοψη S (Εξίσωση 6.10).
$CGain\_trace(T, S)$	κέρδος συμπαγότητας λόγω της αντικατάστασης του μονοπατιού T από τη σύνοψη S (Εξίσωση 6.11).

Πίνακας 6.1: Λίστα συμβόλων για το Κεφάλαιο 6

μία συστάδα που εντοπίστηκε σε κάποια χρονική στιγμή μεταξύ των συστάδων της επόμενης χρονικής στιγμής. Για το σκοπό αυτό εισάγουμε την έννοια της επικάλυψης συστάδων και του ταιριάσματος συστάδων (Ενότητα 6.3.1). Στη συνέχεια, παρουσιάζουμε τις διαφορετικές μεταβολές που μπορεί να υποστεί μία συστάδα, χωρίζοντάς τες σε εξωτερικές (Ενότητα6.3.2.1) και εσωτερικές αλλαγές (Ενότητα 6.3.2.2). Στην Ενότητα 6.3.2.1, παρουσιάζουμε έναν αλγόριθμο για τον εντοπισμό αυτών των μεταβολών.

Με το $MONIC$, ο στόχος μας είναι η δημιουργία ενός πλαισίου παρακολούθησης το οποίο θα είναι ανεξάρτητο από τον τύπο των συστάδων που παρακολουθούνται. Για το σκοπό αυτό, υιοθετούμε τον ορισμό των συστάδων ως σύνολα αντικειμένων (Ορισμός 7), ο οποίος όπως έχουμε ήδη αναφέρει ισχύει για τους διάφορους τύπους συστάδων.

Αργότερα, βέβαια επεκτείνουμε το $MONIC$ στο $MONIC+$ (Ενότητα 6.4) που λαμβάνει επίσης υπόψη και τα ειδικά χαρακτηριστικά κάθε τύπου προτύπων και άρα εξαρτάται από τον τύπο των προτύπων.



Σχήμα 6.3: Ένας δυναμικός πληθυσμός σε δύο χρονικές στιγμές t_1 (επάνω), t_2 (κάτω)

6.3.1 Ταίριασμα συστάδων

Έστω μία συστάδα X που ανακαλύφθηκε την t_i ως μέρος της συσταδοποίησης ζ_i. Μία *μεταβολή συστάδας* είναι μία αλλαγή στη συστάδα αυτή όταν την παρατηρούμε σε κάποια μεταγενέστερη στιγμή t_j $(t_j > t_i)$. Το πρώτο βήμα για τον εντοπισμό

μίας τέτοιας αλλαγής είναι ο εντοπισμός του X στην αντίστοιχη συσταδοποίηση ζ_j, αν υπάρχει ακόμα. Έτσι, ορίζουμε την έννοια της μη συμμετρικής *επικάλυψης* (overlap) και του καλύτερου ταιριάσματος (match) για μία συστάδα, πριν προχωρήσουμε στην παρουσίαση των μεταβολών.

Ορισμός 9 (Επικάλυψη συστάδας) *Έστω ζ_i, ζ_j $(i \neq j)$ είναι δύο συσταδοποιήσεις στις χρονικές στιγμές t_i, t_j, αντίστοιχα και έστω $X \in \zeta_i, Y \in \zeta_j$ είναι δύο συστάδες· Η επικάλυψη της X με την Y είναι το κανονικοποιημένο άθροισμα των βαρών των κοινών τους εγγραφών:*

$$overlap(X,Y) = \frac{\sum_{a \in X \cap Y} age(a, t_j) \times weight(a)}{\sum_{x \in X} age(x, t_j) \times weight(x)} \tag{6.2}$$

Δηλαδή, η επικάλυψη της X με την Y εξαρτάται από το ποσοστό των εγγραφών της X που έχουν επιζήσει στην t_j και ανήκουν στην Y.

Ορίζουμε τώρα για κάθε συστάδα κάποιας στιγμής t_i, το καλύτερο της ταίριασμα σε κάποια μεταγενέστερη στιγμή t_j.

Ορισμός 10 (Ταίριασμα συστάδας) *Έστω X είναι μία συστάδα στη συσταδοποίηση ζ_i τη χρονική στιγμή t_i και Y είναι μία συστάδα στη συσταδοποίηση ζ_j την $t_j > t_i$. Επιπλέον, έστω $\tau \equiv \tau_{match} \in (0.5, 1]$ είναι μία τιμή κατωφλίου. Η Y αποτελεί* ταίριασμα για την X στη ζ_j βάσει του τ, *δηλαδή, $Y = match_\tau(X, \zeta_j)$ αν και μόνο αν:*

1. *η Y έχει τη μέγιστη επικάλυψη με την X ανάμεσα σε όλες τις συστάδες τις ζ_j, δηλαδή, $overlap(X,Y) = \max_{Y' \in \zeta_j}\{overlap(X,Y')\}$ και*
2. *$overlap(X,Y) \geq \tau$.*

Αν δεν υπάρχει τέτοιο $Y \in \zeta_j$, τότε $match_\tau(X, \zeta_j) = \emptyset$.

Από τον Ορισμό 10, η ζ_j μπορεί να περιέχει το πολύ ένα ταίριασμα για κάθε συστάδα στην ζ_i, αν και μία ίδια συστάδα του ζ_j μπορεί να αποτελεί ταίριασμα για παραπάνω από μία συστάδες του ζ_i. Περιορίζουμε το κατώφλι τ στο διάστημα $(0.5, 1]$, προκειμένου να επιβάλλουμε ότι μία συστάδα έχει ταίριασμα μία άλλη συστάδα μόνο αν πάνω από τα μισά μέλη της πρώτης συστάδας έχουν μετακομίσει στη δεύτερη.

Να σημειώσουμε πως σε αυτή την ενότητα ορίσαμε την επικάλυψη/ ομοιότητα μεταξύ δύο συστάδων ως την επικάλυψη των αντικειμένων-μελών τους, υιοθετώντας τον ορισμό των συστάδων ως σύνολα αντικειμένων (Ορισμός 7). Πράγματι, έτσι είναι καθώς ο στόχος μας με το $MONIC$ είναι η δημιουργία ενός πλαισίου εντοπισμού μεταβολών στις συστάδες, το οποίο να είναι ανεξάρτητο από τον τύπο συστάδων. Και όπως έχουμε ήδη αναφέρει η μοντελοποίηση των συστάδων με βάση τα αντικείμενα - μέλη τους ισχύει για τους διάφορους τύπους προτύπων. Αργότερα βέβαια θα δούμε πως η έννοια της επικάλυψης μπορεί να οριστεί ανάλογα με τον τύπο των συστάδων (Ενότητα 6.4).

6.3.2 Μεταβολές συστάδων στο $MONIC$

Στο $MONIC$, μία *μεταβολή* σε μία συγκεκριμένη χρονική στιγμή είναι μία αλλαγή που υπέστη κάποια συστάδα που είχε ανακαλυφθεί σε κάποια προηγούμενη χρονική στιγμή. Μία τέτοια μεταβολή μπορεί να αφορά το πως σχετίζεται η συστάδα με

τις υπόλοιπες συστάδες της συσταδοποίησης στην οποία ανήκει, δηλαδή, να είναι *εξωτερική*, ή μπορεί να αφορά στο περιεχόμενο και τη μορφή της συστάδας, δηλαδή, να είναι *εσωτερική*. Πρώτα ορίζουμε τους διάφορους τύπους μεταβολών και στην συνέχεια προτείνουμε δείκτες για τον εντοπισμό τους.

6.3.2.1 Εντοπισμός εξωτερικών μεταβολών

Οι εξωτερικές μεταβολές που μπορεί να υποστεί μία συστάδα $X \in \zeta_i$ σε σχέση με τη συσταδοποίηση ζ_j ατ $t_j > t_i$ ορίζονται στον Πίνακα 6.2: Μία συστάδα μπορεί να εξαφανιστεί, να διασπαστεί σε πολλαπλές συστάδες, να απορροφηθεί από κάποια άλλη συστάδα ή να επιβιώσει/ επιζήσει, ενώ μπορεί να έχει υποστεί εσωτερικές μεταβολές. Οι εσωτερικές μεταβολές μίας συστάδας παρουσιάζονται στον Πίνακα 6.3: μία συστάδα μπορεί π.χ., να συρρικνωθεί (ή να διογκωθεί) και/ή να γίνει πιο συμπαγής (ή πιο αραιή).

Μία συστάδα $X \in \zeta_i$ *επιβιώνει (survive)* στη συσταδοποίηση ζ_j αν (α) υπάρχει ένα ταίριασμα για αυτή στην ζ_j με βάση το κατώφλι τ και (β) αυτό το ταίριασμα δεν περιλαμβάνει καμία άλλη συστάδα του ζ_i. Αν αυτό το ταίριασμα καλύπτει τουλάχιστον μία ακόμα συστάδα στην ζ_i, τότε η X έχει *απορροφηθεί (absorbed)*. Αν δεν υπάρχει ταίριασμα, τότε μία *διάσπαση (split)* μπορεί να έχει συμβεί: Τα περιεχόμενα της X υπάρχουν σε πάνω από μία συστάδες της ζ_j. Στην περίπτωση αυτή, οι επικαλύψεις δε θα πρέπει να είναι μικρότερες από τ_{split} (προφανώς: $\tau_{split} < \tau$). Επιπλέον, όλες οι συστάδες μαζί θα πρέπει να αποτελούν ταίριασμα για τη X. Αν τίποτα από αυτά δε συμβαίνει, τότε η X έχει *εξαφανιστεί (disappeared)*. Οι *ανερχόμενες (emerging)* συστάδες εντοπίζονται μετά τον εντοπισμό όλων των άλλων εξωτερικών μεταβολών για κάθε συστάδα της ζ_i: είναι οι συστάδες της ζ_j που δεν προέρχονται από κάποια εξωτερική μεταβολή.

Μεταβολή	***Συμβολισμός***	***Δείκτης***
η συστάδα επιβιώνει	$X \rightarrow Y$	$Y = match_\tau(X, \zeta_j)$ και $\nexists Z \in \zeta_i \setminus \{X\} : Y = match_\tau(Z, \zeta_j)$
η συστάδα διασπάται σε πολλαπλές συστάδες	$X \xrightarrow{\subseteq} \{Y_1, \ldots, Y_p\}$	$\forall u = 1 \ldots p : overlap(X, Y_u) \geq \tau_{split}$ και $overlap(X, \cap_{u=1}^{p} Y_u) \geq \tau$ και $(\nexists Y \in \zeta_j \setminus \{Y_1 \ldots, Y_p\} : overlap(X, Y) \geq \tau_{split})$
η συστάδα απορροφάται	$X \xrightarrow{\subseteq} Y$	$Y = match_\tau(X, \zeta_j)$ ΑΝΔ $\exists Z \in \zeta_i \setminus \{X\} : Y = match_\tau(Z, \zeta_j)$
η συστάδα εξαφανίζεται	$X \rightarrow \odot$	καμία από τις παραπάνω περιπτώσεις δεν ισχύει
εμφανίση μίας νέας συστάδας	$\odot \rightarrow X$	

Πίνακας 6.2: Εξωτερικές μεταβολές μίας συστάδας

Στην Εικόνα 6.4 παρουσιάζουμε έναν αλγόριθμο για τον εντοπισμό των αλλαγών, καλούμενο $detectTransitions()$, ο οποίος εντοπίζει τις εξωτερικές μεταβολές των συστάδων την t_i (συσταδοποίηση $\zeta\_i \equiv \zeta_i$ στην εικόνα) σε σχέση με τις συστάδες κάποιας επόμενης στιγμής t_j ($t_j > t_i$) (συσταδοποίηση $\zeta\_j \equiv \zeta_j$ στην εικόνα).

Ο αλγόριθμος υπολογίζει αρχικά έναν πίνακα επικαλύψεων μεταξύ των συστάδων των δύο συσταδοποιήσεων (γραμμή 1). Δεδομένου ότι αυτός ο υπολογισμός

```
detectTransitions()
Input: ζ_i, ζ_j
Output: cluster transitions between ζ_i to ζ_j
BEGIN
1  overlap = computeOverlaps(ζ_i, ζ_j)                          //Matrix of overlaps
2  FOR X ∈ ζ_i
3       splitCandidates = splitUnion = deadList = splitList = absorptionList
        =absorptionSurvivals = absorptionCandidates = ∅;
4       survivalCandidate = NULL;
5       FOR Y ∈ ζ_j
6            Mcell = overlap(X,Y);
7            IF Mcell ≥ τ_match THEN
8                 IF g(X,Y) > g(X,survivalCandidate) THEN
9                      survivalCandidate = Y;
10                ENDIF
11           ELSEIF Mcell ≥ τ_split THEN
12                splitCandidates += Y;
13                splitUnion = splitUnion ∪Y ;
14           ENDIF
15      ENDFOR
16      IF survivalCandidate == NULL OR splitCandidates == ∅
17           THEN deadList += X;                                  //X → ⊙
18      ELSEIF splitCandidates ≠ ∅ THEN
19           IF overlap(X,splitUnion) ≥ τ_match THEN
20                FOR Y ∈ splitCandidates
21                     splitList += (X,Y);
22                ENDFOR                                  //X ⊆→ splitCandidates
23           ELSE deadList += X;                                  //X → ⊙
24           ENDIF
25      ELSE absorptionSurvivals += (X,survivalCandidate);
26      ENDIF
27 ENDFOR
28 FOR Y ∈ ζ_j
29      absorptionCandidates = makeList(absorptionSurvivals,Y);
30      IF cardinality(absorptionCandidates) > 1 THEN
31           FOR X ∈ absorptionCandidates
32             absorptionList +=(X,Y);                            //X ⊆→ Y
33             absorptionSurvivals -= (X,Y);
34           ENDFOR
35      ELSEIF absorptionCandidates == X THEN
36           survivalList +=(X,Y);                                //X → Y
37           absorptionSurvivals -= (X,Y);
38      ENDIF
39 ENDFOR
END
```

Figure 6.4: Εντοπισμός εξωτερικών μεταβολών

είναι ακριβός, τον υπολογίζουμε μία φορά, στην αρχή, και κάθε φορά που χρειαζόμαστε την επικάλυψη δύο συστάδων προσπελαύνουμε το κατάλληλο κελί του πίνα-

κα. Στη συνέχεια, για κάθε συστάδα $X \in \zeta_i$, ο αλγόριθμος κάνει κάτι αρχικοποιήσεις (οι μεταβλητές θα εξηγηθούν στη συνέχεια) και ανακτά την επικάλυψή του με κάθε συστάδα της συσταδοποίησης ζ_j (γραμμή 6). Ο αλγόριθμος ψάχνει πρώτα για συστάδες στην ζ_j που αποτελούν ταίριασμα της X (γραμμές 7–10). Στη γραμμή 8, το καλύτερο ταίριασμα για την X επιλέγεται. Συνεπώς, κάθε συστάδα της ζ_i έχει το πολύ μία υποψήφια συστάδα για επιβίωση. Αν η X δεν έχει καμία, οι συστάδες που έχουν επικάλυψη με αυτή περισσότερο από $\tau_{split} < \tau_{match}$ εντοπίζονται (γραμμές 11–14). Αν δεν υπάρχει καμία, τότε η X θεωρείται ότι έχει εξαφανιστεί (γραμμές 16–17).

Για τον εντοπισμό διασπάσεων συστάδων δημιουργούμε μία λίστα με υποψήφιες συστάδες για διάσπαση (γραμμή 12). Όπως ορίστηκε στον Πίνακα 6.2, αυτές οι συστάδες θα πρέπει όλες μαζί να αποτελούν ένα ταίριασμα για την X. Το όλες μαζί (γραμμή 13) αναφέρεται προς το παρόν στην απλή ένωση των συνόλων των εγγραφών των επιμέρους συστάδων, δηλαδή, τα βάρη δεν λαμβάνονται υπόψη. Ωστόσο, τα βάρη εξακολουθούν να λαμβάνονται υπόψη στον έλεγχο επικάλυψης (γραμμή 19). Αν αυτός ο έλεγχος επιτύχει, η X μαρκάρεται ότι έχει διασπαστεί (γραμμή 21), ειδάλλως μαρκάρεται ότι έχει εξαφανιστεί (γραμμή 23).

Οι περιπτώσεις της απορρόφησης και της επιβίωσης αρχικά αντιμετωπίζονται μαζί: οι συστάδες της ζ_i και οι υποψήφιες συστάδες προς επιβίωση προστίθενται σε μία λίστα από απορροφήσεις και επιβιώσεις (γραμμή 25). Όταν όλες οι συστάδες της ζ_i έχουν επεξεργαστεί, η λίστα αυτή ολοκληρώνεται (γραμμή 27). Τότε, για κάθε συστάδα Y της ζ_j, ο αλγόριθμος εξάγει από τη λίστα αυτή όλες της συστάδες της ζ_i οι οποίες έχουν την Y ως υποψήφια συστάδα επιβίωσης (γραμμή 28). Αν αυτή η υπο-λίστα περιέχει παραπάνω από μία συστάδα, τότε οι συστάδες αυτές έχουν απορροφηθεί από τηνY: Μαρκάρονται λοιπόν (γραμμές 31–32) και απομακρύνονται από την αρχική λίστα (γραμμή 33). Ειδάλλως, αν αυτή η υπο-λίστα αποτελείται από μία μόνο συστάδα τότε η X έχει επιβιώσει στην Y (γραμμή 36). Και πάλι, η αρχική λίστα ενημερώνεται (γραμμή 37).

Μία βελτιστοποίηση πραγματοποιείται στον έλεγχο εντοπισμού διάσπασης (γραμμή 19): Λόγω του ότι θεωρούμε αυστηρή συσταδοποίηση στο πρόβλημά μας, οι συστάδες της ζ_j δεν μπορούν να έχουν κοινά μέλη, ισχύει ότι:

$$\sum_{a \in X \cap (\cup_{u=1}^{p} Y_u)} age(a, t_j) = \sum_{u=1}^{p} \sum_{a \in X \cap Y_u} age(a, t_j)$$

Αυτό μας επιτρέπει να εκτελούμε τον έλεγχο διάσπασης από τις επιμέρους τιμές επικάλυψης του πίνακα επικαλύψεων, χωρίς επιπλέον υπολογισμούς:

$$overlap(X, \cup_{u=1}^{p} Y_u) = \sum_{u=1}^{p} overlap(X, Y_u)$$

Η πολυπλοκότητα του αλγορίθμου Η πολυπλοκότητας περιλαμβάνει το κόστος υπολογισμού του πίνακα επικαλύψεων (ο οποίος υπολογίζεται μία φορά στην αρχή του αλγορίθμου) και το κόστος της διαδικασίας εντοπισμού των μεταβολών.

Όσον αφορά στη διαδικασία εντοπισμού μεταβολών, το κόστος είναι $\mathcal{O}(|\zeta_i| * |\zeta_j|)$. Όσον αφορά στην κατασκευή του πίνακα επικαλύψεων το κόστος είναι $O(|D_i| * |D_j|)$, όπου $D_i(D_j)$ είναι το σύνολο δεδομένων της συσταδοποίησης την $t_i(t_j)$. Συνεπώς, το ολικό κόστος του αλγορίθμου είναι $\mathcal{O}(|\zeta_i| * |\zeta_j| + |D_i| * |D_j|)$.

Στην πράξη, ο αλγόριθμος είναι τετραγωνικός ως προς το μέγεθος του συνόλου δεδομένων που εμπλέκεται στη διαδικασία συσταδοποίησης.

6.3.2.2 Εντοπισμός εσωτερικών μεταβολών

Οι συστάδες που επιζούν μπορεί να υποστούν εσωτερικές αλλαγές Στον Πίνακα 6.3, έχουμε ομαδοποιήσει τις εσωτερικές αλλαγές ως αλλαγές στο μέγεθος, την πυκνότητα και την τοποθεσία. Αν και οι αλλαγές μέσα σε μία ομάδα είναι αμοιβαία αποκλειστικές, οι αλλαγές διαφορετικών ομάδων μπορούν να συνδυαστούν. Για παράδειγμα, μία συστάδα $X \in \zeta_i$ που έχει επιζήσει σε μία συστάδα $Y \in \zeta_j$ μπορεί να γίνει μεγαλύτερη και πιο συμπαγής ταυτόχρονα. Είναι σημαντικό να τονίσουμε εδώ πως υπάρχουν μεταβολές όπως στο μέγεθος που μπορούν να εντοπιστούν κατευθείαν πάνω στα μέλη των συστάδων, ενώ υπάρχουν άλλες μεταβολές, όπως η πυκνότητα και η τοποθεσία, που απαιτούν τον υπολογισμό στατιστικών πάνω στα δεδομένα των συστάδων.

Ομάδα	***Μεταβολή***		***Συμβολισμός***	***Δείκτης***
1.	*Μεγέθους (size)*			
	1α.	συρρίκνωση (shrink)	$X \searrow Y$	$\sum_{x \in X} f(x, t_i) > \sum_{y \in Y} f(y, t_j) + \varepsilon$
	1β.	διόγκωση (expand)	$X \nearrow Y$	$\sum_{y \in Y} f(y, t_j) > \sum_{x \in X} f(x, t_i) + \varepsilon$
2.	*Συμπαγότητας (compactness)*			
	2α.	πιο συμπαγής (more compact)	$X \xrightarrow{\bullet} Y$	$\sigma(Y) < \sigma(X) - \delta$
	2β.	λιγότερη συμπαγής (πιο αραιή) (more diffuse)	$X \xrightarrow{\star} Y$	$\sigma(Y) > \sigma(X) + \delta$
3.	*Θέσης (location shift)*		$X \cdots \rightarrow Y$	I1. $\lvert\mu(X) - \mu(Y)\rvert > \tau_1$ I2. $\lvert\gamma(X) - \gamma(Y)\rvert > \tau_2$ (βλέπε Εξίσωση 6.5)
4.	*Καμία*		$X \leftrightarrow Y$	

Πίνακας 6.3: Εσωτερικές μεταβολές μίας συστάδας

Οι δύο δείκτες για τον εντοπισμό μεταβολών στο μέγεθος συγκρίνουν τα σύνολα δεδομένων των X και Y, λαμβάνοντας συνεπώς υπόψη τα βάρη των δεδομένων λόγω της γήρανσης. Ωστόσο τα βάρη χρησιμοποιούνται διαφορετικά σε σχέση με τον υπολογισμό της επικάλυψης. Συγκεκριμένα, ενώ για την επικάλυψη χρησιμοποιούνται τα βάρη την t_j, για την μεταβολή στο μέγεθος θεωρούμε τα βάρη των εγγραφών του X στην πραγματική στιγμή t_i. Αυτό είναι λογικό δεδομένου ότι για τη μεταβολή στο μέγεθος θα πρέπει να θεωρήσουμε τη σημαντικότητα των επιμέρους εγγραφών των συστάδων τη στιγμή t_i σε σχέση με τη στιγμή t_j.

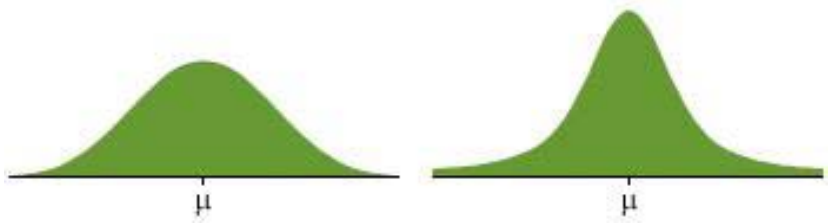
Η μεταβολή πυκνότητας δεν μπορεί να εντοπιστεί παρατηρώντας κατευθείαν τα δεδομένα, αλλά απαιτεί στατιστικές τιμές πάνω στα δεδομένα. Ο δείκτης σ που εμφανίζεται στον Πίνακα 6.3 είναι η τυπική απόκλιση:

$$\sigma(X) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu(X))^2} \tag{6.3}$$

όπου $\mu(X)$ είναι η μέση τιμή X.:Το κατώφλι δ χρησιμοποιείται για να εμποδίσουμε μικρές αλλαγές να θεωρηθούν ως αλλαγές πυκνότητας. Άλλα στατιστικά μέτρα πάνω στα δεδομένα μπορούν να χρησιμοποιηθούν, όπως η κύρτωση:

$$kurtosis(X) = \frac{\frac{1}{card(X)} \sum_{x \in X} (x - \mu(X))^4}{\left(\frac{1}{card(X)} \sum_{x \in X} (x - \mu(X))^2 \right)^2} - 3 \tag{6.4}$$

Στην Εικόνα 6.5 [68], παρουσιάζουμε ένα παράδειγμα δύο κατανομών με διαφορετικές τιμές κύρτωσης, η δεξιά κατανομή έχει μεγαλύτερη κύρτωση σε σχέση με την αριστερή.

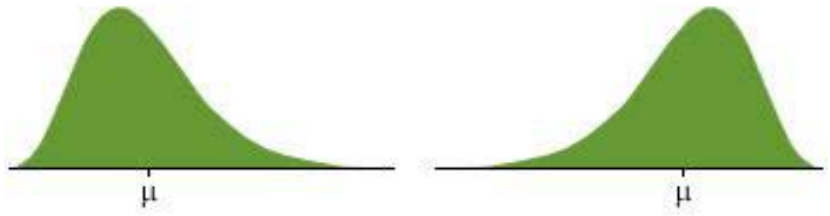


Σχήμα 6.5: Παράδειγμα κύρτωσης

Στην ειδική περίπτωση ενός στατικού μετρικού χώρου, οι αλλαγές του Πίνακα 6.3 μπορούν να εντοπιστούν μελετώντας τις τοπολογικές ιδιότητες των συστάδων. Σε ένα μετρικό χώρο, μία συστάδα μπορεί επίσης να μετατοπιστεί. Ακόμα και αν δεν υπάρχει μετρικός χώρος μπορούμε να εντοπίσουμε αλλαγές στην τοποθεσία ως μετατοπίσεις στην κατανομή των δεδομένων: ο δείκτης I1 εντοπίζει αλλαγές στο μέσο όρο $\mu(X)$, ενώ ο δείκτης I2 εντοπίζει αλλαγές στην ασυμμετρία $\gamma(X)$:

$$\gamma(X) = \frac{\frac{1}{card(X)} \sum_{x \in X} (x - \mu(X))^3}{\left(\frac{1}{card(X)} \sum_{x \in X} (x - \mu(X))^2 \right)^{\frac{3}{2}}} \tag{6.5}$$

Στην Εικόνα 6.6 [69], παρουσιάζουμε ένα παράδειγμα δύο κατανομών με διαφορετικές τιμές ασυμμετρίας, η αριστερή έχει αρνητική ασυμμετρία, ενώ η δεξιά κατανομή έχει θετική ασυμμετρία.



Σχήμα 6.6: Παράδειγμα ασυμμετρίας

6.4 Το πλαίσιο $MONIC+$ για διάφορους τύπους συστάδων

Το πλαίσιο $MONIC$ (Ενότητα 6.3) εντοπίζει μεταβολές πάνω σε συστάδες που αναπαρίστανται ως σύνολα από αντικείμενα (Ορισμός 7). Στην ενότητα αυτή, επεκτείνουμε το $MONIC$ στο $MONIC+$ που επιπλέον καλύπτει και τα ειδικά χαρακτηριστικά που έχει κάθε τύπος συστάδας, επιτρέποντάς μας να εντοπίζουμε μεταβολές συστάδων ανάλογα με τον τύπο των συστάδων.

Πρώτα ορίζουμε την έννοια της επικάλυψης/ ομοιότητας για τους διαφορετικούς τύπους συστάδων (Ενότητα 6.4.1). Στη συνέχεια (Ενότητα 6.4.2), περιγράφουμε τις διαφορετικές μεταβολές που υποστηρίζει κάθε τύπος προτύπων, όπως επίσης και δείκτες για τον εντοπισμός τους.

6.4.1 Ταίριασμα συστάδων για διαφορετικούς τύπους συστάδων

Στην Ενότητα 6.3.1, εισάγαμε την έννοια της επικάλυψης και του ταιριάσματος για μία συστάδα προκειμένου να αποφασίσουμε κατά πόσο μία συστάδα που βρέθηκε την t_i εξακολουθεί να υπάρχει και την επόμενη χρονική στιγμή t_j. Ωστόσο, ο ορισμός της έννοιας της επικάλυψης σε εκείνη την ενότητα αναφέρεται σε συστάδες που ορίζονται ως σύνολα αντικειμένων, δηλαδή, Τύπος B1 (Ορισμός 7). Στην ενότητα αυτή, ορίζουμε την έννοια της επικάλυψης για διάφορους τύπους συστάδων. Συγκεκριμένα, ορίζουμε μία γενική συνάρτηση επικάλυψης, την οποία αρχικοποιούμε κατάλληλα ανάλογα με τον τύπο συστάδων.

Ορισμός 11 (Επικάλυψης συστάδων) *Έστω ζ_i είναι η συσταδοποίηση την t_i και ζ_j η συσταδοποίηση την $t_j, j \neq i$. Ορίζουμε τη συνάρτηση $overlap()$ που υπολογίζει την ομοιότητα ή επικάλυψη της συστάδας $X \in \zeta_i$ σε σχέση με μία συστάδα $Y \in \zeta_j$ ως μία τιμή στο διάστημα $[0..1]$ έτσι ώστε (ι) η τιμή 1 να δηλώνει μέγιστη επικάλυψη και η τιμή 0 να δηλώνει έλλειψη επικάλυψης και (ιι) να ισχύει ότι $\sum_{Y \in \zeta_j} overlap(X,Y) \leq 1$.*

Η επικάλυψη ορίζεται μη συμμετρικά. Στη συνέχεια, καθορίζουμε τη συνάρτηση $overlap()$ για κάθε τύπο προτύπων.

Ορισμός 12 (Επικάλυψη για συστάδες τύπου A) *Έστω ζ_i, ζ_j είναι δύο συσταδοποιήσεις συστάδων τύπου A, τις χρονικές στιγμές $t_i < t_j$, αντίστοιχα. Για δύο συστάδες $X \in \zeta_i$ και $Y \in \zeta_j$, η επικάλυψη της X στην Y είναι η κανονικοποιημένη τομή των περιοχών τους:*

$$overlap(X,Y) = \frac{area(X) \cap area(Y)}{area(X)} \tag{6.6}$$

Ορισμός 13 (Επικάλυψη για συστάδες τύπου B1)

Όπως στον Ορισμό 9.

Ορισμός 14 (Επικάλυψη για συστάδες τύπου B2) *Έστω ζ_i, ζ_j είναι δύο συσταδοποιήσεις συστάδων τύπου A, τις χρονικές στιγμές $t_i < t_j$, αντίστοιχα. Για*

δύο συστάδες $X \in \zeta_i$ και $Y \in \zeta_j$, η επικάλυψη της X στην Y ορίζεται με βάση την εγγύτητα των κέντρων τους:

$$overlap(X,Y) = \begin{cases} 1 - \frac{|\mu(X)-\mu(Y)|}{\sigma(X)} & , \quad |\mu(X)-\mu(Y)| \leq \sigma(X) \\ 0 & , \quad otherwise \end{cases} \tag{6.7}$$

Για κάθε συστάδα που εντοπίζεται την t_i, μπορούμε να βρούμε το καλύτερό της ταίριασμα κάποια επόμενη στιγμή t_j χρησιμοποιώντας τη συνάρτηση *cluster match*.

6.4.2 Εντοπισμός μεταβολών με βάση τον τύπο συστάδων

Ο εντοπισμός των εξωτερικών μεταβολών στο $MONIC+$ είναι όπως στο $MONIC$ (βλέπε Ενότητα 6.3), αλλά μερικά βήματα πρέπει να υλοποιηθούν διαφορετικά ανάλογα με τον τύπο της συστάδας.

Οι μεταβολές που παρατηρούνται για κάθε τύπο συστάδας παρουσιάζονται στον Πίνακα 6.4. Όλες οι εξωτερικές και εσωτερικές μεταβολές μπορούν να εντοπιστούν για συστάδες σε ένα μετρικό χώρο (Τύπου Α). Για συστάδες τύπου Β1, οι μεταβολές συμπαγότητας και τοποθεσίας μπορούν δεν να εντοπιστούν κατευθείαν, γιατί τέτοιες έννοιες δεν ορίζονται κατευθείαν στα δεδομένα. Ωστόσο, όταν κάποιος έχει μια πιο περιγραφική (*intensional*) αναπαράστασης της συστάδας, οι μεταβολές αυτές μπορούν να εντοπιστούν ως αλλαγές στην κατανομή πυκνότητας, αναφερόμαστε στον τύπο αυτό ως Τύπος συστάδας Β1+Β2. Με την ίδια λογική, ο ορισμός μίας συστάδας τύπου Β2 δεν επιτρέπει τον εντοπισμό διασπάσεων και απορροφήσεων, ο οποίες μπορούν να εντοπιστούν μελετώντας τα μέλη των συστάδων (Τύπος Β1+Β2).

	Μεταβολές			
	Εξωτερικές	***Εσωτερικές μεταβολές***		
Τύπος συστάδας		*Μέγεθος*	*Συμπαγότητα*	*Τοποθεσία*
A. μετρικός χώρος	Ναι	Ναι	Ναι	Ναι
μη μετρικός χώρος				
B1. εξωτερικές	Ναι	Ναι	Όχι	Όχι
B2. εσωτερικές	επιβίωση	Ναι	Ναι	Ναι
B1+B2.	Ναι	Ναι	Ναι	Ναι

Πίνακας 6.4: Δυνατές μεταβολές για κάθε τύπο συστάδας

Δείκτες μεταβολών για συστάδες τύπου Α. Έστω ζ_i, ζ_j είναι οι συσταδοποιήσεις τις χρονικές στιγμές $t_i < t_j$ και έστω $X \in \zeta_i$ η συστάδα που παρατηρούμε. Οι δείκτες μεταβολών που προτείνονται στον Πίνακα 6.5 χρησιμοποιούν τον ορισμό της επικάλυψης για συστάδες τύπου Α (Ορισμός 12) και τον αντίστοιχο ορισμό του ταιριάσματος συστάδων (Ορισμός 10).

Οι εξωτερικές μεταβολές των συστάδων εντοπίζονται βρίσκοντας την περιοχή επικάλυψης μεταξύ της συστάδας X και κάθε συστάδας της ζ_j. Για τον εντοπισμό διάσπασης, προσαρμόζουμε τον έλεγχο διάσπασης του Αλγορίθμου 6.4. Πιο

συγκεκριμένα, υπολογίζουμε την επικάλυψη μεταξύ της περιοχής της X και της περιοχής όλων των υποψηφίων $Y_1, Y_2, \ldots, Y_p$. Δεδομένου ότι οι υποψήφιες συστάδες δεν επικαλύπτονται, χρησιμοποιούμε την ακόλουθη εξίσωση για να κάνουμε τον έλεγχο διάσπασης:

$$area(X) \cap area(\cup_{u=1}^{p} Y_u) = \sum_{u=1}^{p} area(X) \cap area(Y_u) \tag{6.8}$$

Βήμα	***Μεταβολή***	***Δείκτης***
1	Επιβίωση ή Απορρόφηση	$\exists Y \in \zeta_j : \frac{area(X) \cap area(Y)}{area(X)} \geq \tau$
2	$X \xrightarrow{\subseteq} Y$	$\exists Z \in \zeta_i \setminus \{X\} : \frac{area(Z) \cap area(Y)}{area(Z)} \geq \tau$
3	$X \rightarrow Y$	$\nexists Z \in \zeta_i \setminus \{X\} : \frac{area(Z) \cap area(Y)}{area(Z)} \geq \tau$
4	$X \xrightarrow{\subseteq} \{Y_1, \ldots, Y_p\}$	$\exists Y_1, \ldots, Y_p \in \zeta_1 :$ $(\forall Y_u : \frac{area(X) \cap area(Y_u)}{area(X)} \geq \tau_{split}) \wedge \frac{area(X) \cap area(\cup_{u=1}^{p} Y_u)}{area(X)} \geq \tau$
5	$X \rightarrow \odot$	απορρέουν από τα παραπάνω
6	$X \nearrow \searrow Y$	B1 δείκτες & B2 δείκτες
7	$X \xrightarrow{\bullet\star} Y$	με βάση τη γεωμετρία & B2 δείκτες
8	$X \cdots \rightarrow Y$	με βάση τη γεωμετρία & B2 δείκτες

Πίνακας 6.5: Δείκτες μεταβολών για συστάδες τύπου A

Ο εντοπισμός εσωτερικών μεταβολών αντιστοιχεί στην παρακολούθηση της τροχιάς της συστάδας σε ένα στατικό μετρικό χώρο. Στον Πίνακα 6.6, προτείνουμε δείκτες για σφαιρικές συστάδες, όπως παράγονται από τους αλγορίθμους π.χ., K-Means και K-Medoids. Μπορούμε επίσης να χρησιμοποιήσουμε και δείκτες μεταβολών από αυτούς που ισχύουν για τους τύπους συστάδων B1 και B2.

Μεταβολή	***Δείκτης***
$X \cdots \rightarrow Y$	$\frac{d(center(X), center(Y))}{min\{radius(X), radius(Y)\}} \geq \tau_{location}$
$X \xrightarrow{\bullet} Y$	$avg_{x \in X}(d(x, center(X)) > avg_{y \in Y}(d(y, center(Y)) + \varepsilon$
$X \xrightarrow{\star} Y$	$avg_{y \in Y}(d(y, center(Y)) > avg_{x \in X}(d(x, center(X)) + \varepsilon$

Πίνακας 6.6: Δείκτες μεταβολών για σφαιρικές συστάδες

Το πρώτο *heuristic* στον Πίνακα 6.6 εντοπίζει μεταβολές τοποθεσίας ελέγχοντας αν η απόσταση μεταξύ των κέντρων των συστάδων υπερβαίνει ένα κατώφλι $\tau_{location}$.Το δεύτερο *heuristic* δηλώνει πως μία συστάδα έχει γίνει πιο συμπαγής αν η μέση απόσταση από το κέντρο της συστάδας στην παλιά συστάδα είναι μεγαλύτερη από ότι στην καινούρια συστάδα, θεωρώντας και ένα μικρό κατώφλι ε. Το τρίτο *heuristic* για συστάδες που γίνονται πιο αραιές είναι το αντίθετο του δεύτερου.

Δείκτες μεταβολών για συστάδες τύπου B1. Οι δείκτες μεταβολών για συστάδες τύπου B1 έχουν ήδη περιγραφεί στην Ενότητα 6.3, όπου παρουσιάσαμε το πλαίσιο $MONIC$. Έτσι, δεν τους επαναλαμβάνουμε εδώ. Απλά να σημειώσουμε πως όλες οι εξωτερικές μεταβολές μπορούν να εντοπιστούν για συστάδες

τύπου B1, ενώ από τις εσωτερικές μεταβολές μόνο οι μεταβολές μεγέθους μπορούν να εντοπιστούν εξετάζοντας κατευθείαν τα δεδομένα-μέλη των συστάδων.

Δείκτες μεταβολών για συστάδες τύπου B2 Θεωρούμε πάλι μία συστάδα $X \in \zeta_i$. Για τον εντοπισμό μεταβολών μεγέθους, χρησιμοποιούμε τα *heuristic* για συστάδες τύπου B1 (Πίνακας 6.3). Για τις υπόλοιπες μεταβολές που μπορούμε να εντοπίσουμε(Πίνακας 6.4), χρησιμοποιούμε τους δείκτες του Πίνακα 6.7. Ο πρώτος δείκτης λέει πως μία συστάδα επιβιώνει αν υπάρχει ταίριασμα για αυτή, με βάση το κατώφλι $\tau \in (0.5, 1]$ (Ορισμός 10).

Βήμα	***Μεταβολή***	***Δείκτης***
1	$X \rightarrow Y$	$\exists Y \in \zeta_j : 1 - \frac{\lvert\mu(X)-\mu(Y)\rvert}{\sigma(X)} \geq \tau$
2	$X \rightarrow \odot$	άρνηση του προηγούμενου
3	$X \nearrow \searrow Y$	B1 δείκτες (Πίνακας 6.3)
4	$X \cdots \rightarrow Y$	η1. $\lvert\mu(X) - \mu(Y)\rvert > \tau_{h1}$ η2. $\lvert\gamma(X) - \gamma(Y)\rvert > \tau_{h2}$ (βλέπε Εξίσωση 6.5)
5	Συμπαγότητα $X \xrightarrow{\bullet} Y$ $X \xrightarrow{\star} Y$	 $\sigma(Y) < \sigma(X) + \varepsilon$ (Εξίσωση 6.3) $\sigma(X) < \sigma(Y) + \varepsilon$ (Εξίσωση 6.3)

Πίνακας 6.7: Δείκτες μεταβολών για συστάδες τύπου B2

Ο εντοπισμός μιας μεταβολής απορρόφησης για την $X \in \zeta_i$ σημαίνει την εύρεση μίας συστάδας $Y \in \zeta_j$ που περιέχει τα $X, Z \in \zeta_i$. Παρομοίως, η εύρεση μίας μεταβολής διάσπασης αντιστοιχεί στην εύρεση συστάδων που περιέχουν υποσύνολα της X. Ωστόσο, αυτό συνεπάγεται ότι χειριζόμαστε τις συστάδες σαν σύνολα δεδομένων (Τύπος B1). Έτσι, θεωρούμε μόνο επιβιώσεις και εξαφανίσεις για τις συστάδες τύπου B2.

Για το εντοπισμό μεταβολών συμπαγότητας, χρησιμοποιούμε τη διαφορά των τυπικών αποκλίσεων των συστάδων X, Y. Για τον εντοπισμό μεταβολών τοποθεσίας, χρησιμοποιούμε δύο *heuristics* που εντοπίζουν διαφορετικού τύπου μετατόπισης στην τοποθεσία: ο $h1$ εντοπίζει μετατόπιση στην τιμή του μέσου, ενώ ο $h2$ εντοπίζει αλλαγές στην ασυμμετρία (*skewness*) $\gamma()$ (Εξίσωση 6.5).

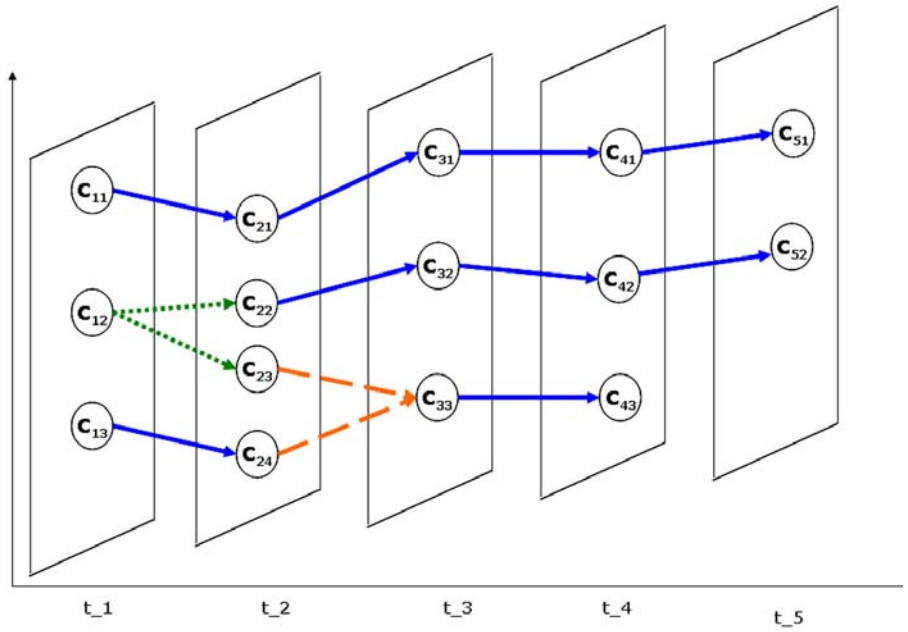
6.5 Η ιστορία της εξέλιξης (*Evolution Graph*)

Το πλαίσιο $MONIC$ (και συνεπώς και το $MONIC+$), εντοπίζει μεταβολές μεταξύ συσταδοποιήσεων που ανακαλύφθηκαν μεταξύ διαδοχικών χρονικών στιγμών. Καθώς η περίοδος παρατήρησης αυξάνεται, συναθροίζονται όλο και περισσότερες συστάδες και μεταβολές μεταξύ των συστάδων.

Μοντελοποιούμε την ιστορία της εξέλιξης του πληθυσμού με μία δομή γράφου, την επονομαζόμενη *Evolution Graph* $EG \equiv G(V, E)$, που εκτείνεται σε ολόκληρη την περίοδο παρατήρησης του πληθυσμού (n χρονικές στιγμές). Το σύνολο των κόμβων V αντιστοιχεί στο σύνολο των συστάδων που εμφανίστηκαν κατά τη διάρκεια αυτής της χρονικής στιγμής: $V = \{\zeta_1, \ldots, \zeta_n\}$, όπου κάθε συσταδοποίηση $\zeta_i \in V$ περιέχει τις συστάδες που ανακαλύφθηκαν την αντίστοιχη χρονική στιγμή

t_i, δηλαδή, $\zeta_i = \{c_1, c_2, \ldots, c_{|\zeta_i|}\}$. Το σύνολο των ακμών E περιέχει τις μεταβολές των συστάδων: $\forall e = (c, c') \in E$ όπου υπάρχει μία χρονική στιγμή $t_i, 1 \leq i < n$ τέτοια ώστε $c \in \zeta_i$ και $c' \in \zeta_{i+1}$. Με αυτές τις προδιαγραφές του *Evolution Graph*, οι ακμές συνδέουν κόμβους/συστάδες που βρίσκονται σε διαδοχικές χρονικές στιγμές.

Ένα παράδειγμα ενός *Evolution Graph* ις παρουσιάζεται στην Εικόνα 6.7, όπου το χρώμα και το είδος των γραμμών/ακμών υποδεικνύει τη σημασιολογία των μεταβολών. Μία *στικτή (πράσινη)* ακμή υποδηλώνει μία διάσπαση της αρχικής συστάδας σε πολλαπλές συστάδες (π.χ., $c_{12} \xrightarrow{\subseteq} \{c_{22}, c_{23}\}$). Μία *διακεκομένη (πορτοκαλί)* ακμή περιγράφει μία απορρόφιση. Οι συστάδες απορροφώνται από την τελική συστάδα (π.χ., $\{c_{23}, c_{24}\} \xrightarrow{\subseteq} c_{33}$). Μία *συνεχής (μπλε)* ακμή υποδηλώνει μία επιβίωση. Μία συστάδα επιβιώνει σε μία άλλη συστάδα (π.χ., $c_{11} \rightarrow c_{21}$). Κόμβοι χωρίς εισερχόμενες ακμές αντιστοιχούν σε πρωτοεμφανιζόμενες συστάδες (π.χ., συστάδα c_{11}), ενώ κόμβοι χωρίς εξερχόμενες ακμές αντιστοιχούν σε συστάδες που εξαφανίζονται (π.χ., συστάδα c_{43}).



Σχήμα 6.7: Παράδειγμα ενός Γράφου Εξέλιξης (EG)

6.5.1 Το μοντέλο του *Evolution Graph*

Πρώτα περιγράφουμε τη σημασιολογία των κόμβων (Ενότητα 6.5.1.1) και στη συνέχεια τη σημασιολογία των ακμών του γράφου (Σεχτιον 6.5.1.2), τα οποία χρησιμοποιούμε στη συνέχεια για να αναπτύξουμε το μοντέλο του *Evolution Graph*.

6.5.1.1 Σημασιολογία των κόμβων του γράφου

Ένας κόμβος $c \in V$ αναπαριστά μία συστάδα που εντοπίζεται την χρονική στιγμή t_i, δηλαδή, και ανήκει σε μία συσταδοποίηση ζ_i. Αναθέτουμε σε κάθε κόμβο του γράφου μία ετικέτα $\widehat{c}$ που συνοψίζει τα μέλη της αντίστοιχης συστάδας σε κάποια περιγραφική/ διαισθητική μορφή. Πολλές πολύπλοκες συνοπτικές αναπαραστάσεις συστάδων υπάρχουν στη βιβλιογραφία, για παράδειγμα οι συνόψεις συστάδων του [24], οι μικρο-συστάδες του [2] και τα (droplets) για δεδομένα κειμένου [3]. Επιλέγουμε δύο απλές αναπαραστάσεις: τα κέντρα των συστάδων (cluster centroids) για συστάδες που ορίζονται πάνω σε αυθαίρετα αριθμητικά δεδομένα και τις ετικέτες συστάδων για συστάδες που ορίζονται πάνω σε δεδομένα κειμένου.

Ορισμός 15 (Ετικέτα με βάση τον κεντροειδή) *Έστω c μία συστάδα σε έναν m-διάστατο χώρο αριθμητικών γνωρισμάτων. Ο κεντροειδής της c είναι ένα διάνυσμα μέσων τιμών $\widehat{c} :=< \mu_1 \dots \mu_m >$, όπου μ_l είναι η μέση τιμή των τιμών των δεδομένων επί της l^{th}-διάστασης, $1 \leq l \leq m$.*

Θεωρούμε ότι οι τιμές των δεδομένων είναι κανονικοποιημένες στο διάστημα [0..1].

Για συστάδες κειμένων, προτιμάμε τη μοντελοποίηση με βάση τις πιο συχνές λέξεις κλειδιά. Θεωρούμε ότι σε ένα βήμα προεπεξεργασίας αφαιρούνται οι συνηθισμένες λέξεις και τα stopwords, έτσι ώστε οι λέξεις που απομένουν να είναι εκείνες που τελικώς χαρακτηρίζουν το περιεχόμενο της συστάδας.

Ορισμός 16 (Ετικέτα με βάση τις λέξεις κλειδιά) Έστω c μία συστάδα κειμένων, όπου κάθε κείμενο $d_i \in c$ είναι ένα διάνυσμα στον χώρο γνωρισμάτων των λέξεων κλειδιών $\{k_1, \dots, k_m\}$. Η ετικέτα της συστάδας ορίζεται ως $\widehat{c} :=< w_{k_1}, \dots, w_{k_m} >$, όπου w_{k_l} είναι (α) η συχνότητα της l^{th}-λέξης μέσα στη c, εάν η συχνότητα ξεπερνά ένα όριο b, ή (β) μηδέν, διαφορετικά.

6.5.1.2 Σημασιολογία των ακμών του γράφου

Μία ακμή $e = (c, c') \in E$ υποδηλώνει ότι μία συστάδα $c \in \zeta_i$ έχει εξελιχθεί σε μία άλλη συστάδα $c' \in \zeta_{i+1}$ της επόμενης χρονικής στιγμής. Με την εξέλιξη εννοούμε ότι ανάμεσα στις συστάδες του ζ_{i+1}, η συστάδα c' είναι η πιο όμοια με την συστάδα c. Η συνέχεια αυτή περικλείει διαφορετικές περιπτώσεις, π.χ., η c' μπορεί να είναι πολύ μεγαλύτερη από τη c ή μπορεί να έχει απορροφήσει τη c όπως και άλλες συστάδες του ζ_i, ή τη c μπορεί να έχει διασπαστεί σε πολλαπλές μικρότερες συστάδες στο ζ_{i+1} ή ο κεντροειδής της συστάδας να έχει μετατοπισθεί σε διαφορετική τοποθεσία από το t_i στο t_{i+1}. Σχεδιάζουμε τη σημασιολογία της εξέλιξης των συστάδων σύμφωνα με το πλαίσιο $MONIC$ (βλέπε Ενότητα 6.3).

Πιο συγκεκριμένα, μία ακμή ενώνει το c και το c' στον *Evolution Graph*, εάν υπάρχει κάποια εξωτερική μεταβολή στο c (επιβίωση, διάσπαση, απορρόφηση) η οποία να εξελίσσεται στη c'. Κάθε ακμή $e = (c, c') \in E$ περιέχει και μία ετικέτα $e.ExtTransition$ που περιγράφει τον (εξωτερικό) τύπο μεταβολής (δηλαδή, επιβίωση, διάσπαση, απορρόφηση)μεταξύ της $c \in \zeta_i$ και της $c' \in \zeta_{i+1}$. Εάν μία συστάδα του ζ_i δεν έχει εξερχόμενες ακμές, τότε έχει *εξαφανισθεί*. Εάν μία συστάδα του ζ_{i+1} δεν έχει καμία εισερχόμενη ακμή, τότε αποτελεί μία *νέα συστάδα*. Στην περίπτωση των εσωτερικών μεταβολών, η e περιέχει επίσης μία ετικέτα $e.IntTransition$ που περιγράφει τις εσωτερικές μεταβολές που μία συστάδα περιέχει και τις τιμές αυτών, π.χ., {(συρρίκνωση μεγέθους, 20%) (περισσότερο συμπαγής, 10%)}.

6.5.2 Η κατασκευή του *Evolution Graph*

Ο*Evolution Graph* δημιουργείται σιγά σιγά καθώς έρχονται τα αποτελέσματα των συσταδοποιήσεων $t_1, \ldots, t_n$. Όταν μία νέα συσταδοποίηση ζ_i φτάνει την $t_i, i > 1$, το $MONIC$ (ή το $MONIC+$) εφαρμόζεται στην προηγούμενη συσταδοποίηση ζ_{i-1} και στην τρέχουσα συσταδοποίηση ζ_i: τυχόν μεταβολές μεταξύ των συστάδων των ζ_{i-1}, ζ_i εντοπίζονται και μία ακμή προστίθεται στον *Evolution Graph* για κάθε μεταβολή που εντοπίζεται.

Ο αλγόριθμος δημιουργίας του *Evolution Graph* παρατίθεται στην Εικόνα 6.8.

---

buildEG()
Output: $EG = G(V, E)$
BEGIN
1. WHILE a new clustering ζ_i arrives at t_i BEGIN
2. EG.addNodes($\zeta\_i.nodes$); *//add ζ_i clusters in EG*
3. IF ($i > 1$) THEN
   $j = i - 1$; *//for notation*
4. E_ji=detectTransitions(ζ_j, ζ_i); *//detect transitions*
5. EG.addEdges(E_ji); *//add transition edges*
6. EG.updateNodes($\zeta\_j.nodes$); *//remove redundant information from ζ_j*
7. ENDIF;
8. END;
9. return EG;
END

Figure 6.8: Ο αλγόριθμος δημιουργίας του *Evolution Graph*

Ο αλγόριθμος προσθέτει τους κόμβους και τις ακμές της συσταδοποίησης ζ_i (ζ_i στην εικόνα αυτή) στον *Evolution Graph* και τους αναθέτει ετικέτες (γραμμή 2). Αν το ζ_i είναι η πρώτη συσταδοποίηση του *Evolution Graph*, δεν χρειάζονται περαιτέρω ενέργειες. Διαφορετικά (γραμμές 3–7), οι μεταβολές της ζ_i σε σχέση με την προηγούμενη συσταδοποίηση ζ_{i-1}, που είναι ήδη στον EG, εντοπίζονται και προστίθενται στο γράφο. Για τον εντοπισμό των μεταβολών, καλείται ο αλγόριθμος του $MONIC$ (βλέπε Αλγόριθμος 6.4) (γραμμή 4). Αν το $MONIC$ βρει επιβιώσεις, διασπάσεις ή απορροφήσεις στην ζ_i, οι αντίστοιχες ακμές προστίθενται και εμπλουτίζονται με πληροφορία σχετικά με τη μεταβολή (γραμμή 5). Αν κάποιες από τις μεταβολές που εντοπίστηκαν είναι επιβιώσεις, οι αντίστοιχες ακμές του EG εμπλουτίζονται επίσης και με πληροφορία σχετικά με τον τύπο των (πιθανών) εσωτερικών μεταβολών (μέγεθος, συμπαγότητα, τοποθεσία).

Το $MONIC$ χρησιμοποιεί τα περιεχόμενα των συστάδων για τον εντοπισμό των μεταβολών. Συνεπώς, κρατάμε την πληροφορία αυτή μέχρι την επόμενη χρονική στιγμή μόνο. Οι κόμβοι που εισέρχονται την t_{i-1} ενημερώνονται (γραμμή 6): η ετικέτα κάθε κόμβου υπολογίζεται και αποθηκεύεται ενώ τα ίδια τα δεδομένα απομακρύνονται.

Πολυπλοκότητα αλγορίθμου Μελετάμε τώρα την πολυπλοκότητα του προτεινόμενου αλγορίθμου. Η προσθήκη μίας νέας συστάδας ζ_i στον *Evolution Graph* επιφέρει τα ακόλουθα κόστη: ι) τον υπολογισμό των ετικετών για τις επιμέρους συστάδες της συγκεκριμένης συσταδοποίησης και, ιι) τον εντοπισμό

των μεταβολών μεταξύ της συγκεκριμένης συσταδοποίησης (ζ_i) και της συσταδοποίησης της προηγούμενης χρονικής στιγμής (ζ_j). Έστω $cost(c)$ είναι το κόστος υπολογισμού ετικέτας μίας συστάδας c. Τότε, το κόστος υπολογισμού ετικετών για όλες τις συστάδες του ζ_i είναι: $|\zeta_i| * cost(c)$. Το κόστος εντοπισμού μεταβολών μεταξύ των ζ_i, ζ_j είναι το κόστος του $MONIC$: $O(|\zeta_i| * |\zeta_j| + |D_i| * |D_j|)$ (βλέπε Ενότητα 6.3.2.1). Συνεπώς, το κόστος για την προσθήκη της ζ_i στον *Evolution Graph*είναι:

$$addCost = O(|\zeta_i| * |\zeta_j| + |D_i| * |D_j| + |\zeta_i| * cost(c))$$

Συνεπώς, αν θεωρήσουμε μία περίοδο μελέτης από n χρονικές στιγμές, το *κόστος κατασκευής* γίνεται:

$$EGbuildCost = (n-1) * addCost$$

Υπάρχει επίσης και το κόστος αποθήκευσης του *Evolution Graph*, το οποίο αναφέρεται στο χώρο που απαιτείται για την αποθήκευση του *Evolution Graph*. Για κάποιο t_i, θα πρέπει να αποθηκεύσουμε τις ετικέτες των συστάδων της αντίστοιχης συσταδοποίησης ζ_i και τις μεταβολές συστάδων μεταξύ αυτής της συσταδοποίησης και της συσταδοποίησης της προηγούμενης χρονικής στιγμής, t_j. Αν $|centroid|$ είναι το κόστος αποθήκευσης της ετικέτας μίας συστάδας, τότε το κόστος αποθήκευσης για τις ετικέτες των συστάδων της ζ_i είναι $|\zeta_i| * |centroid|$. Αν $|edge|$ είναι το κόστος αποθήκευσης μιας ακμής, τότε το κόστος αποθήκευσης των μεταβολών μεταξύ των ζ_i, ζ_j είναι: $|\zeta_i| * |\zeta_j| * |edge|$.

Για μία περίοδο από n χρονικές στιγμές, το κόστος αποθήκευσης γίνεται:

$$EGstorageCost = \sum_{i=1...n} |\zeta_i| \times |centroid| + \sum_{i=1...n-1, j=i+1} |\zeta_i| \times |\zeta_j| \times |edge|$$

Να σημειώσουμε ότι κατά την κατασκευή του *Evolution Graph*, θα πρέπει επίσης να αποθηκεύσουμε και τα περιεχόμενα της πιο πρόσφατης συσταδοποίησης, ζ_j. Αυτό είναι απαραίτητο γιατί προκειμένου να εντοπίσουμε τις μεταβολές της επόμενης συσταδοποίησης ζ_i σε σχέση με την ζ_j, θα πρέπει να έχουμε στη διάθεσή μας τα περιεχόμενα της ζ_j. Το κόστος αυτό είναι $O(|D_j|)$, όπου D_j είναι το σύνολο δεδομένων στο οποίο στηρίζεται η συσταδοποίηση ζ_j.

6.5.3 Το σύνολο μονοπατιών του *Evolution Graph*

Έστω $t_i, 1 \leq i \leq n$, είναι μία χρονική στιγμή της περιόδου παρατήρησης. Μπορεί να υπάρχουν συστάδες που σχηματίζονται για πρώτη φορά στην t_i, δηλαδή δεν αποτελούν επιβιώσεις από την προηγούμενη χρονική στιγμή t_{i-1}, τις ονομάζουμε *αναδυόμενες συστάδες (emerged clusters)*. Βάσει του ορισμού, οι αναδυόμενες συστάδες μπορεί να είναι είτε συστάδες που εμφανίστηκαν (appearance) για πρώτη φορά την t_i ή τα αποτελέσματα κάποιας μεταβολής διάσπασης ή απορρόφησης από την t_{i-1}.

Για μία αναδυόμενη συστάδα, σχηματίζουμε το *μονοπάτι της συστάδας (cluster trace* ως εξής:

Ορισμός 17 (Μονοπάτι συστάδας (Cluster Trace» *Έστω c είναι μία αναδυόμενη συστάδα την t_i. Η ακολουθία $\prec c_i \cdot c_{i+1} \cdot \cdot c_k \succ$ των συστάδων που ανακαλύπτονται τις $t_i, t_{i+1}, \ldots, t_k$, $1 < k \leq n$ είναι το μονοπάτι της c, $trace(c)$, αν $c_i \equiv c$ και για*

κάθε $c_j, j > 1$ *υπάρχει μία ακμή* $e = (c_{j-1}, c_j)$ *έτσι ώστε* $e.transition = survival$. *Επιπλέον, η τελευταία συστάδα αυτής της ακολουθίας, δηλαδή η* c_k, *δεν επιβιώνει στην επόμενη χρονική στιγμή* t_{k+1}.

Η έννοια του μονοπατιού μιας συστάδας μας επιτρέπει να βρούμε τη διαδρομή μίας συστάδας μέσα στον εξελισσόμενο πληθυσμό από την πρώτη φορά που η συστάδα σχηματίστηκε μέχρι τη στιγμή που διαλύθηκε. Συμβολίζουμε με $\mathcal{T}_{EG}$ το *σύνολο μονοπατιών (traceset) του EG*. Το *traceset* $\mathcal{T}_{EG}$ του παραδείγματός μας (Εικόνα 6.7) αποτελείται από τα ακόλουθα μονοπάτια:

- τραςε $\prec c_{11}c_{21}c_{31}c_{41}c_{51} \succ$, που δείχνει ότι η συστάδα c_{11} που αναθύθηκε την t_1 επιβίωσε για πέντε χρονικές στιγμές,
- τραςε $\prec c_{22}c_{32}c_{42}c_{52} \succ$ της συστάδας c_{22}, μίας από τις συστάδες στις οποίες διασπάστηκε η c_{12} και
- δύο μονοπάτια 2 κόμβων $\prec c_{13}c_{24} \succ$ και $\prec c_{33}c_{43} \succ$.

Οι άλλες συστάδες c_{12}, c_{23}, c_{24} υπάρχουν μόνο για μία χρονική στιγμή και συνεπώς δεν δημιουργούν μονοπάτια. Έτσι:

$$\mathcal{T}_{EG} = \{\prec c_{11}c_{21}c_{31}c_{41}c_{51} \succ, \prec c_{22}c_{32}c_{42}c_{52} \succ, \prec c_{13}c_{24} \succ, \prec c_{33}c_{43} \succ\}$$

Για την εξαγωγή του *traceset* του *Evolution Graph* μπορεί κανείς νά διατρέξει το γράφο ξεκινώντας από την πρώτη στιγμή της περιόδου παρατήρησης και δημιουργώντας τα μονοπάτια συστάδων με βάση τον Ορισμό 17.

6.5.4 Η αξιοποίηση του *Evolution Graph*

Ο *Evolution Graph* περιέχει την όλη ιστορία των μεταβολών των συστάδων, συνεπώς μπορεί να αξιοποιηθεί για τον εντοπισμό και την κατανόηση των αλλαγών στον πρωταρχικό δυναμικό πληθυσμό. Παρακάτω παρουσιάζουμε δύο τέτοιες δυνατότητες αξιοποίησης: η πρώτη επιτρέπει στο χρήστη να κερδίσει κατανόηση σχετικά με την ευστάθεια του πληθυσμού κατά μήκος του χρόνου εξέλιξης μελετώντας το χρόνο ζωής των συστάδων και των συσταδοποιήσεων (Ενότητα 6.5.4.1), ενώ η δεύτερη παρέχει στο χρήση δυνατότητες επερωτήσεων πάνω στην ιστορία της εξέλιξης του πληθυσμού (Ενότητα 6.5.4.2).

6.5.4.1 Χρόνος ζωής συστάδων και συσταδοποιήσεων

Ο εντοπισμός μεταβολών στις συστάδες προσφέρει γνώση σχετικά με την εξέλιξη μεμονωμένων συστάδων αλλά και ολόκληρης της συσταδοποίησης. Διαισθητικά, αν οι περισσότερες από τις συστάδες μίας συσταδοποίησης επιζούν σε επόμενη περίοδο, τότε ο πληθυσμός είναι σχετικά στατικός. Αν όμως οι συστάδες υπόκεινται σε συχνές μεταβολές τύπου π.χ., *split* τότε ο πληθυσμός είναι δυναμικός και η εκάστοτε συσταδοποίηση δεν περιγράφει καλά δεδομένα επόμενων στιγμών.

Ορίζουμε (α) το χρόνο ζωής μίας συστάδας (cluster lifetime) και (β) το χρόνο ζωής μίας συσταδοποίησης (clustering lifetime) ώστε να αποκτήσουμε γνώση σχετικά με τη σταθερότητα του εξελισσόμενου πληθυσμού κατά μήκος της περιόδου παρατήρησης.

Ορισμός 18 (Χρόνος ζωής συστάδας - Cluster lifetime) *Έστω C είναι μία συστάδα και t_i είναι η χρονική στιγμή που εμφανίστηκε για πρώτη φορά ως μέρος της συσταδοποίησης ζ_i). Ο χρόνος ζωής της C είναι το πλήθος των χρονικών στιγμών στις οποίες η C έχει επιζήσει. Ορίζουμε (ι)* αυστηρό χρόνο ζωής *$lifetimeS$ ως το πλήθος των διαδοχικών στιγμών που έχει επιζήσει χωρίς εσωτερικές μεταβολές, (ιι)* χρόνο ζωής με εσωτερικές αλλαγές *$lifetimeI$ όπου λαμβάνονται υπόψη και οι περιπτώσεις που έχει επιζήσει με εσωτερικές αλλαγές και (ιιι)* χρόνο ζωής με *absorptions $lifetimeA$ που επιπλέον λαμβάνει υπόψη και τα absorptions της C.*

Βάσει του ορισμού αυτού, ο χρόνος ζωής μίας συστάδας είναι τουλάχιστον 1 καθώς εμφανίζεται σίγουρα μία χρονική στιγμή. Υπολογίζουμε το χρόνο ζωής των συστάδων ως εξής: Ξεκινάμε με την ζ_n και θέτουμε ίσο με 1 το χρόνο ζωής όλων των συστάδων της. Σε κάποια προγενέστερη στιγμή t_i, ο αυστηρός χρόνος ζωής της συστάδας X είναι 1 αν η X δεν επιβίωσε στην t_{i+1}. Αν υπάρχει κάποια $Y \in \zeta_{i+1}$ με $X \leftrightarrow Y$, τότε $lifetimeS(X) = lifetimeS(Y) + 1$. Αν υπάρχει κάποια $Y \in \zeta_{i+1}$ με $X \rightarrow Y$, τότε ο χρόνος ζωής με εσωτερικές μεταβολές της X είναι: $lifetimeI(X) = lifetimeI(Y) + 1$. Αν υπάρχει κάποια $Y \in \zeta_{i+1}$ έτσι ώστε είτε $X \rightarrow Y$ ή $X \xrightarrow{\subseteq} Y$, τότε ο χρόνος ζωής με απορροφήσεις της X είναι $lifetimeA(X) = lifetimeA(Y) + 1$.

Σημειώστε επίσης ότι $lifetimeI(c) = |trace(c)|$, δηλαδή ο χρόνος ζωής με εσωτερικές μεταβολές για μία συστάδα c ισούται με το μήκος του μονοπατιού (trace) της συστάδας (βλέπε επίσης τον Ορισμό 17).

Το αντίστοιχο του χρόνου ζωής μίας συστάδας, είναι ο χρόνος ζωής μίας συσταδοποίησης.

Ορισμός 19 (Χρόνος ζωής συσταδοποίησης -Clustering Lifetime) *Έστω ζ είναι μία συσταδοποίηση. Ο χρόνος ζωής της $L(\zeta)$ είναι το μεσαίο (median) $lifetimeA$ των επιμέρους συστάδων της:*

$$L(\zeta) = median_{C \in \zeta}\{lifetimeA(C)\}$$

Στον ορισμό αυτό, χρησιμοποιούμε τον πιο χαλαρό ορισμό του χρόνου ζωής μιας συστάδας, αυτόν δηλαδή που επιτρέπει εσωτερικές αλλαγές και *absorptions*. Για να αποτρέψουμε την κυριαρχία λίγων συστάδων με μικρό ή μεγάλο χρόνο ζωής χρησιμοποιούμε την έννοια του μεσαίου (*median*) χρόνου αντί του μέσου (*mean*).

Ο χρόνος ζωής μίας συσταδοποίησης είναι μια μακροπρόθεσμη ιδιότητα. Στις περισσότερες περιπτώσεις οι συσταδοποιήσεις έχουν μικρό χρόνο ζωής ακόμα και αν μερικές από τις συστάδες τους επιζούν σε παραπάνω χρονικές στιγμές. Ορίζουμε συνεπώς μία βραχυπρόθεσμη έκδοση του χρόνου ζωής μίας συσταδοποίησης, με βάση το ποσοστό των συστάδων που επιζούν ή απορροφούνται (*absorb*) στην επόμενη χρονική στιγμή.

Ορισμός 20 (Ρυθμός επιβίωσης - Survival Ratio) *Έστω ζ_i είναι η συσταδοποίηση την t_i για $i = 1, \ldots, n-1$. Ο ρυθμός επιβίωσης είναι το ποσοστό των συστάδων της ζ_i που επιζούν (πιθανόν με εσωτερικές αλλαγές) στην ζ_{i+1}:*

$$survivalRatio(\zeta_i) = \frac{|\{X \in \zeta_i | \exists Y \in \zeta_{i+1} : X \rightarrow Y\}|}{|\zeta_i|}$$

Ο ρυθμός επιβίωσης λαμβάνει υπόψη μόνο τις συστάδες που έχουν επιζήσει στην επόμενη χρονική στιγμή. Χαλαρώνουμε τον ορισμό λαμβάνοντας επίσης υπόψη και τις συστάδες που έχουν απορροφηθεί.

Ορισμός 21 (Ρυθμός απορρόφησης - Absorption Ratio) *Ο ρυθμός απορρόφησης του ζ_i είναι το ποσοστό των συστάδων του που απορροφούνται από τις συστάδες της ζ_{i+1}:*

$$absorptionRatio(\zeta_i) = \frac{|\{X \in \zeta_i | \exists Y \in \zeta_{i+1} : X \xrightarrow{\subseteq} Y\}|}{|\zeta_i|}$$

Θεωρώντας τόσο το ρυθμό επιβίωσης όσο και το ρυθμό απορρόφησης, προκύπτει ο λεγόμενος ρυθμός $passforward$, ο οποίος αποτελεί μέτρο της " αντοχής " του πληθυσμού την επόμενη χρονική στιγμή

Ορισμός 22 (Ρυθμός $Passforward$**)** *Ο ρυθμός $passforward$ του ζ_i είναι το ποσοστό των συστάδων της που επέζησε ή απορροφήθηκε από συστάδες της ζ_{i+1}.*

$$passforwardRatio(\zeta_i) = survivalRatio(\zeta_i) + absorptionRatio(\zeta_i)$$

Ο ρυθμός $passforward$ αναπαριστά το βαθμό στον οποίο μία συσταδοποίηση περιγράφει/ χαρακτηρίζει τα δεδομένα της επόμενης χρονικής στιγμής. Αν είναι χαμηλός τότε η συσταδοποίηση έχει μικρό χρόνο ζωής, ακόμα και αν κάποιες από τις συστάδες της μπορεί να επιζήσουν περισσότερο.

6.5.4.2 Επερωτήσεις στον *Evolution Graph*

Ο *Evolution Graph* περιέχει έναν πλούτο πληροφορίας σε σχέση με την εξέλιξη του πληθυσμού που μελετάμε. Διαφορετικά είδη ερωτημάτων μπορούν να οριστούν πάνω στον *Evolution Graph* προκειμένου να διευκολύνουν τον τελικό χρήστη να αποκτήσει καλύτερη κατανόηση του εξελισσόμενου πληθυσμού.

Μερικά ενδεικτικά ερωτήματα είναι ως ακολούθως:

- **Forward History Ερωτήματα:** *Πως εξελίσσεται η συστάδα X μετά την χρονική στιγμή t_i;*
  Answer sketch: Ξεκίνα από την συστάδα X και ακολούθα τις εξερχόμενες ακμές της έως ότου οι απόγονοί της να εξαφανιστούν.

- **Backward History Ερωτήματα:** *Πως αναδύθηκε η συστάδα X;*
  Answer sketch: Ξεκινά από τη συστάδα X και ακολούθα τις εισερχόμενες ακμές της έως ότου οι πρόγονοί της να εμφανιστούν για πρώτη φορά.

- **Ερωτήματα Ομοιότητας:** *Ποιες άλλες συστάδες έχουν παρόμοια ιστορία μεταβολών ($forward$ ή $backward$ ή και τα δύο) με την συστάδα Ξ;*
  Answer sketch: Χρησιμοποίησε την ιστορία της συστάδας X και δες αν η ακολουθία μεταβολών που έχει υποστεί η X μοιάζει με την ακολουθία μεταβολών που έχει υποστεί κάποια άλλη συστάδα του πληθυσμού.

- **Ερωτήματα επιρροής:** *Ποιες συστάδες και σε ποιες χρονικές στιγμές έχουν περισσότερο επηρεάσει την X ως προς της μορφή και το περιεχόμενό της;*

Answer sketch: Ανέθεσε έναν παράγοντα σημαντικότητας σε κάθε συστάδα Y που συμμετέχει στην ιστορία της X με βάση παράγοντες όπως: η επικάλυψη της Y σε σχέση με τις εξερχόμενες συστάδες/ κόμβους κατά μήκος της ιστορίας της X, η απόσταση της Y (π.χ., βάσει του πλήθους των ενδιάμεσων ακμών) από την X κ.ο.κ. Διέταξε τις συστάδες με βάση το βαθμό σημαντικότητάς τους.

6.6 Το πλαίσιο $FINGERPRINT$ για την συμπίεση της εξέλιξης των συστάδων

Ο *Evolution Graph* περιέχει όλη την ιστορία της εξέλιξης του πληθυσμού και επιτρέπει τη μελέτη των μεταβολών των συστάδων στην πορεία του χρόνου. Ωστόσο, καθώς η περίοδος παρατήρησης μεγαλώνει, ο *Evolution Graph* διογκώνεται σημαντικά και αξιοποίησή του με κάποιο χειρωνακτικό τρόπο είναι δύσκολη. Ποιος θα μπορούσε να καταλάβει για παράδειγμα ένα γράφο που αποτελείται από 1000 κόμβους και τις μεταβολές μεταξύ τους· Σίγουρα, αυτός ο όγκος πληροφορίας μπορεί δύσκολα να γίνει κατανοητός από τον τελικό χρήστη και συνεπώς, υπάρχει ανάγκη για συμπίεση του γράφου σε κάποια πιο συμπαγή μορφή που όμως να διατηρεί την πληροφορία του αρχικού γράφου.

Για το σκοπό αυτό, εκμεταλλευόμαστε το γεγονός ότι ο γράφος περιέχει επικαλύψεις. Πιο συγκεκριμένα, ο γράφος περιέχει πληροφορίες σχετικά με συστάδες που μεταβάλλονται αλλά και σχετικά με συστάδες που παραμένουν ίδιες ή μεταβάλλονται ελαφρώς. Έτσι, συνοψίζουμε τον *Evolution Graph* με τέτοιο τρόπο που να μην υπάρχουν επικαλύψεις αλλά και να παραμένουν οι σημαντικές μεταβολές. Για το σκοπό αυτό, συνοψίζουμε τα μονοπάτια συστάδων (cluster traces), δηλαδή, ακολουθίες επιβιώσεων, σε μία συμπυκνωμένη μορφή που την αποκαλούμε *fingerprint*. Οι συνόψεις όλων των μονοπατιών συστάδων αποτελούν το $FINGERPRINT$ του*Evolution Graph*.

Επιλέγουμε να συνοψίσουμε μονοπάτια συστάδων, δηλαδή, επιβιώσεις συστάδων, επειδή μία επιβίωση από μία συστάδα σε μία άλλη δείχνει ότι η αρχική συστάδα είναι κάπως παρόμοια με την τελική συστάδα. Αντιθέτως, κάποια διάσπαση ή απορρόφηση συστάδας υποδηλώνει ότι η αρχική συστάδα υπέστη κάποια σημαντική αλλαγή και μία τέτοια αλλαγή πρέπει να την αναφέρουμε στο χρήστη αντί να την αποσιωπήσουμε. Το ίδιο ισχύει με τις νεο-εμφανιζόμενες συστάδες - θα πρέπει να τις αναφέρουμε στον τελικό χρήστη γιατί αποκαλύπτουν τη δημιουργία νέων ομάδων στον πληθυσμό.

Πρώτα μοντελοποιούμε την έννοια της σύνοψης για ένα μονοπάτι συστάδας και μετράμε την απώλεια πληροφορίας και το κέρδος συμπαγότητας σε σχέση με το αρχικό μονοπάτι της συστάδας, λόγω της διαδικασίας συμπίεσης (Ενότητα 6.6.1). Στη συνέχεια, προτείνουμε διαφορετικές μεθόδους συμπίεσης των μονοπατιών λαμβάνοντας υπόψη την απώλεια πληροφορίας και το κέρδος συμπαγότητας που προκύπτουν.

6.6.1 Σύνοψη μονοπατιού (*trace summary*)

Η διαδικασία συμπίεσης/σύνοψης εφαρμόζεται πάνω στα μονοπάτια των συστάδων (cluster trace), τα οποία ορίστηκαν στον Ορισμό 17. Διατρέχουμε κάθε μονοπάτι T και εντοπίζουμε τις συστάδες/κόμβους που μπορούν να απομακρυνθούν: είναι οι συστάδες/κόμβοι που μπορούν να αντικατασταθούν από ένα μικρότερο αριθμό

παραγόμενων κόμβων, τους οποίους ονομάζουμε "εικονικά κέντρα" (virtual centers) και ορίζονται ως ακολούθως:

Ορισμός 23 (Εικονικό κέντρο (Virtual Center» *Έστω $\prec c_1 \dots c_m \succ$ είναι το μονοπάτι μια αναδυόμενης συστάδας c, $trace(c)$, και έστω $X = \prec c_j \dots c_{j+k} \succ$ είναι ένα υπο-μονοπάτι αυτού (subtrace), δηλαδή, μία υπο-ακολουθία από γειτονικούς κόμβους του αρχικού μονοπατιού ($k \leq m-1$, $j \geq 1$). Ορίζουμε το "εικονικό κέντρο" της X, $\widehat{X}$, ως έναν παραγώμενο κόμβο ο οποίος αποτελείται από τους μέσους όρους των ετικετών των κόμβων στο X:*

$$\widehat{X}[i] = \frac{1}{|X|} \sum_{c_i \in X} \widehat{c}[i]$$

όπου $\cdot[i]$ είναι η i διάσταση και το $\widehat{c}$ συμβολίζει την ετικέτα της συστάδας c. Χρησιμοποιούμε το συμβολισμό $c \mapsto \widehat{X}$ για να δείξουμε ότι η συστάδα $c \in X$ έχει αντιστοιχιθεί στο εικονικό κέντρο $\widehat{X}$.

Αν οι ετικέτες ορίζονται με βάση τον *centroid* (Ορισμός 15), το $\widehat{X}$ είναι το κέντρο των *centroids* των συστάδων στο X. Αν οι ετικέτες ορίζονται με βάση τις λέξεις - κλειδιά (Ορισμός 16), τότε το $\widehat{X}$ περιέχει τις μέσες συχνότητες των λέξεων - κλειδιά στο X.

Αφού εισάγαμε την έννοια του εικονικού κέντρου για ένα *subtrace*, ορίζουμε την έννοια της σύνοψης για ένα *trace*: αποτελείται από μία ακολουθία κόμβων όπου κάθε κόμβος είναι είτε ένας πραγματικός κόμβος του αρχικού *trace* είτε ένα εικονικό κέντρο που συνοψίζει κάποιο *subtrace* του αρχικού *trace*.

Ορισμός 24 (Σύνοψη μονοπατιού (Trace Summary» *Έστω $T = \prec c_1 \dots c_m \succ$ είναι ένα trace. Μία ακολουθία $S = \prec a_1 \dots a_k \succ$ αποτελεί μία σύνοψη (summary) του T αν και μόνο αν (a) $k \leq m$ και (β) για κάθε $c_i \in T$ υπάρχει ένα $a_j \in S$ έτσι ώστε είτε $c_i = a_j$ ή $c_i \mapsto a_j$ (δηλαδή, το c_i ανήκει σε κάποιο subtrace του T το οποίο έχει συνοψιστεί στο εικονικό κέντρο a_j).*

Υπάρχουν διάφοροι τρόποι συμπίεσης ενός *trace*, όπου κάθε τρόπος αντιστοιχεί σε μία διαφορετική τμηματοποίηση του *trace* σε *subtraces* με αποτέλεσμα να δημιουργούνται διαφορετικά εικονικά κέντρα. Ενδιαφερόμαστε για συμπιέσεις που επιτυγχάνουν υψηλό κέρδος συμπαγότητας ενώ διατηρούν χαμηλή την απώλεια πληροφορίας. Οι στόχοι αυτοί περιγράφονται μέσω συναρτήσεων που μετρούν το κέρδος συμπαγότητας και την απώλεια πληροφορίας και οι οποίες περιγράφονται παρακάτω.

Η αντικατάσταση της *subtrace* X από ένα εικονικό κέντρο $\widehat{X}$ οδηγεί σε *κέρδος συμπαγότητας*, καθώς λιγότεροι κόμβοι αποθηκεύονται, αλλά επίσης και σε *απώλεια πληροφορίας*, καθώς οι αρχικές συστάδες αντικαθίστανται από εικονικά κέντρα.

Μοντελοποιούμε την απώλεια πληροφορίας (information loss) κάθε αρχικής συστάδας $c \in X$ με βάση την απόστασή της από το εικονικό κέντρο $\widehat{X}$ στο οποίο ανατίθεται μετά τη συμπίεση:

$$ILoss\_cluster(c, \widehat{X}) = dist(\widehat{c}, \widehat{X}) \tag{6.9}$$

όπου $dist(\widehat{c}, \widehat{X})$ είναι η απόσταση της ετικέτας της αρχικής συστάδας $\widehat{c}$ από το εικονικό κέντρο $\widehat{X}$.

Η απώλεια πληροφορίας κάθε συστάδας/ κόμβου ενός *trace* συνυπολογίζεται στην συνολική απώλεια πληροφορίας του *trace* ως ακολούθως:

Ορισμός 25 *Έστω T είναι μία trace και S μία σύνοψη αυτής. Η απώλεια πληροφορίας (information loss) της T σε σχέση με την S ορίζεται ως:*

$$ILoss\_trace(T,S) = \sum_{c \in T} ILoss\_cluster(c, a_c) \tag{6.10}$$

όπου το $a_c \in S$ αντιστοιχεί είτε στο εικονικό κέντρο στο οποίο ανατέθηκε η c μετά τη σύνοψη είτε στην ίδια τη συστάδα c. Στη δεύτερη περίπτωση, $ILoss\_cluster(c, a_c) = 0$.

Το κέρδος συμπαγότητας (compactness gain) του T σε σχέση με το S ορίζεται ως η μείωση στο πλήθος των συστάδων και των ακμών που χρειάζεται να αποθηκευτούν:

$$\begin{aligned} CGain\_trace(T,S) &= \frac{(|T|-|S|)+(|T|-1-(|S|-1))}{|T|+|T|-1} \\ &= \frac{2\times(|T|-|S|)}{2\times|T|-1)} \approx \frac{|T|-|S|}{|T|} \end{aligned} \tag{6.11}$$

όπου $|T|$ είναι ο αριθμός των κόμβων της T και $|T|-1$ είναι ο αριθμός των ακμών μεταξύ αυτών των κόμβων (παρόμοια για το S).

Στη συνέχεια, ορίζουμε το *fingerprint* ενός *trace* ως μία σύνοψη, τα εικονικά κέντρα της οποίας είναι κοντά στις ετικέτες των αρχικών συστάδων, με βάση ένα άνω όριο απόστασης δ, προκειμένου η απώλεια πληροφορίας που προκύπτει λόγω της αντικατάστασης μιας συστάδας από ένα εικονικό κέντρο να παραμένει μικρή.

Ορισμός 26 (Το *Fingerprint* ενός μονοπατιού (Trace fingerprint» *Έστω T είναι ένα trace και S μία σύνοψη αυτού. Το S αποτελεί fingerprint για το T αν και μόνο αν:*

($C1$) *Για κάθε κόμβο $c \in X$ που αντικαθίσταται από ένα εικονικό κέντρο $a \in S$ ισχύει ότι $dist(\widehat{c}, a) \leq \delta$ και*

($C2$) *για κάθε $(sub)trace \prec c_1 \cdot \ldots \cdot c_k \succ$ του T που έχει συμπιεστεί σε ένα εικονικό κέντρο a ισχύει ότι $\forall i = 1l, dots, k-1 : dist(\widehat{c_i}, \widehat{c_{i+1}}) \leq \delta$.*

Με βάση τον ορισμό αυτό, το S αποτελεί *fingerprint* για το T αν έχει διαμερίσει το T σε *subtraces* από συστάδες που είναι παρόμοιες μεταξύ τους (συνθήκη $C2$) και κάθε τέτοιο *subtrace* έχει ένα εικονικό κέντρο που είναι κοντά σε όλες τις αρχικές συστάδες (συνθήκη $\tilde{}1$).

Από τη στιγμή που τα *traces* συμπιέζονται σε *fingerprints*, ο *Evolution Graph* μπορεί επίσης να συμπιεστεί, προκαλώντας απώλεια πληροφορίας και κέρδος συμπαγότητας στο επίπεδο του γράφου.

Ορισμός 27 *Έστω EG είναι ένας Evolution Graph και $\mathcal{T}_{EG}$ είναι το σύνολο των traces του (traceset). Για κάθε trace $T \in \mathcal{T}_{EG}$, έστω S_T είναι το fingerprint του (Ορισμός 26) και έστω δ είναι το άνω όριο στην απόσταση των κεντροειδών. Το σύνολο $\mathcal{S}_{EG} := \{S_T | T \in \mathcal{T}_{EG}\}$ είναι το fingerprint του Evolution Graphκαι*

προκαλεί ένα κέρδος συμπαγότητας CG και μία απώλεια πληροφορίας IL που ορίζονται ως εξής:

$$CG(EG, \mathcal{S}_{EG}) = \sum_{T} CGain\_trace(T, S_T) \quad (6.12)$$

$$IL(EG, \mathcal{S}_{EG}) = \sum_{T} ILoss\_trace(T, S_T) \quad (6.13)$$

Στη συνέχεια περιγράφουμε τον αλγόριθμο *Batch FINGERPRINT* που δημιουργεί το *fingerprint* όλου του *Evolution Graph*τμηματοποιώντας τα *traces* ώστε να δημιουργηθούν τα *fingerprints* τους. Ο αλγόριθμος αυτός προϋποθέτει ότι ο *Evolution Graph*έχει πρώτα κατασκευαστεί και αποθηκευτεί ολόκληρος (batch ή offline αλγόριθμος). Στη συνέχεια, παρουσιάζουμε τον αλγόριθμο *Incremental FINGERPRINT*, ο οποίος δημιουργεί τα *fingerprints* των *traces* αυξητικά/σταδιακά καθώς νέες συστάδες εντοπίζονται (online ή incremental αλγόριθμος). Στην περίπτωση αυτή, το *fingerprint* της εξέλιξης χτίζεται κατευθείαν, χωρίς να χρειάζεται να έχει κατασκευαστεί πρώτα ο *Evolution Graph*.

6.6.2 *Batch* συμπίεση του *Evolution Graph*

Ο αλγόριθμος *Batch FINGERPRINT* συμπιέζει τον *Evolution Graph EG* εντοπίζοντας τα *traces* του και δημιουργώντας ένα *fingerprint* για κάθε *trace*. Στη συνέχεια, αντικαθιστά τα αρχικά *traces* στον *EG* με τα *fingerprints* τους. Ο αλγόριθμος *Batch FINGERPRINT* ικανοποιεί τις δύο συνθήκες του Ορισμού 26 εφαρμόζοντας σε κάθε (*sub*)*trace* T δύο heuristics:

- *Heuristic A:* Αν το T περιέχει γειτονικούς κόμβους που απέχουν μεταξύ τους παραπάνω από δ, τότε το ζεύγος των γειτονικών κόμβων c, c' με τη μέγιστη απόσταση εντοπίζεται και το T διαμερίζεται στα τμήματα T_1, T_2 έτσι ώστε το c να είναι ο τελευταίος κόμβος στην T_1 και ο c' να είναι ο πρώτος κόμβος στην T_2.

- *Heuristic B:* Αν το T ικανοποιεί τη συνθήκη *(C2)* αλλά περιέχει κόμβους που απέχουν παραπάνω από δ από το εικονικό κέντρο, τότε το T διασπάται ως ακολούθως: Ο κόμβος c που έχει τη μεγαλύτερη απόσταση από το εικονικό κέντρο του T, $vcenter(T)$, εντοπίζεται και το T διαμερίζεται στα T_1, T_2 έτσι ώστε ο c να είναι ο τελευταίος κόμβος της T_1 και ο διάδοχός του c' να είναι ο πρώτος κόμβος της T_2.

Το *Heuristic A* αντιμετωπίζει τυχόν παραβιάσεις της συνθήκης *(C2)* και το *Heuristic B* αντιμετωπίζει τυχόν παραβιάσεις της συνθήκης *(C1)* για τα (*sub*)*traces* που ήδη ικανοποιούν το *(C2)*. Παρουσιάζουμε τον αλγόριθμο στην Εικόνα 6.9.

Ο αλγόριθμος *Batch FINGERPRINT* δημιουργεί ένα *fingerprint* του *Evolution Graph* διατρέχοντας το γράφο, εξάγοντας τα *traces* του (γραμμή 1, συνθήκη *C2*) και συμπιέζοντας καθένα από αυτά (γραμμή 4). Τα παραγώμενα *fingerprints* των *traces* προστίθενται στον *FEG* (γραμμή 5). Αυτή η λειτουργία περιλαμβάνει την προσθήκη του *trace fingerprint* στο γράφο και την ανακατεύθυνση των εισερχόμενων/εξερχόμενων ακμών του αρχικού *trace* στα άκρα του αντίστοιχου *fingerprint*.

Ο αλγόριθμος *Batch FINGERPRINT* καλεί τη συνάρτηση *summarize_HeuristicA* που αναδρομικά σπάει το *trace* σε *subtraces* με βάση τον *Heuristic A* έως ότου η

$Batch\ \ FINGERPRINT(EG)$
Input: the $Evolution\ Graph\tilde{}EG$
Output: FEG, a fingerprint of EG
1. traverse the EG and extract its traces into $\mathcal{T}$;
2. $FEG = \emptyset$;
3. for each trace $T \in \mathcal{T}$ do
4. $FT = summarize\_HeuristicA(T)$;
5. $FEG.addTrace(FT)$;;
6. end-for
7. return FEG;

$summarize\_HeuristicA(T)$
Input: a trace T
Output: a fingerprint of the trace
1. if $|T| == 1$ then return T;
2. if C_b is not satisfied then
3. find $c \in T$ such that
$\forall(y,z) \in T\_1 : dist(y,z) < dist(c, c\_next)$ and
$\forall(y,z) \in T\_2 : dist(y,z) < dist(c, c\_next)$;
//(c, c_next) is the most dissimilar pair of consecutive nodes in T
4. split T into $T\_1 =\prec c\_1, \ldots, c \succ$ and $T\_2 =\prec c\_next, \ldots, c\_k \succ$;
5. $FT\_1 = summarize\_HeuristicA(T\_1)$;
6. $FT\_2 = summarize\_HeuristicA(T\_2)$;
7. return $\prec FT\_1 \cdot FT\_2 \succ$;
8. else return $summarize\_HeuristicB(T)$;
9. endif

$summarize\_HeuristicB(T)$
Input: a trace T
Output: a fingerprint of the trace
1. $v = vcenter(T)$;
2. if $\forall y \in T : dist(y,v) < \delta$ then //Condition $C1$
3. return v;
4. else
5. find $c \in T$ such that $dist(c,v) = \max\{dist(y,v) | y \in T\}$;
6. split T into $T\_1 =\prec c\_1, \ldots, c \succ$ and $T\_2 =\prec c\_next, \ldots, c\_k \succ$;
7. $FT\_1 = summarize\_HeuristicB(T\_1)$;
8. $FT\_2 = summarize\_HeuristicB(T\_2)$;
9. return $\prec FT\_1 \cdot FT\_2 \succ$;
10. endif

Figure 6.9: Ο αλγόριθμος $Batch\ \ FINGERPRINT$ για την $offline$ συμπίεση του $Evolution\ Graph$

$(C2)$ να ικανοποιηθεί. Αν το $trace$ αποτελείται από ένα μόνο κόμβο, αυτός ο κόμβος επιστρέφεται (γραμμή 1). Ειδάλλως, ελέγχουμε κατά πόσο το $trace$ περιέχει κόμβους των οποίων οι ετικέτες απέχουν περισσότερο από το κατώφλι δ (γραμμή 2).Αν η $(C2)$ ικανοποιείται, τότε καλείται η συνάρτηση $summarize\_HeuristicB$ (γραμμή 8): Ελέγχει αν ισχύει η συνθήκη $(C1)$ και επιστρέφει το $fingerprint$ της $(sub)trace$ εισόδου. Αν η $(C2)$ παραβιάζεται, το $trace$ διαμερίζεται με βάση το $Heuristic\ A$ (γραμμές 3,4) και το $summarize\_HeuristicA$ καλείται για κάθε

επιμέρους τμήμα (γραμμές 5, 6). Τέλος, οι συμπιεσμένες $(sub)traces$ συνενώνονται (γραμμή 7) και επιστρέφονται.

Η συνάρτηση $summarize\_HeuristicB$ δουλεύει με παρόμοιο τρόπο. Δέχεται ως είσοδο ένα $(sub)trace$ T που αποτελείται από παραπάνω από έναν κόμβους και ικανοποιεί τη συνθήκη $(C2)$. Δημιουργεί τα εικονικά κέντρα για το T με βάση τον Ορισμό 23 και στη συνέχεια ελέγχει τη συνθήκη $(C1)$ συγκρίνοντας την απόσταση μεταξύ του εικονικού κέντρου και κάθε κόμβου από το κατώφλι δ (γραμμή 2). Αν η απόσταση δεν υπερβαίνει το δ, το εικονικό κέντρο επιστρέφεται (γραμμή 3). Ειδάλλως, το T διασπάται στον κόμβο που έχει τη μεγαλύτερη απόσταση από το εικονικό κέντρο, με βάση το *Heuristic B* (γραμμές 5, 6). Το $summarize\_HeuristicB$ καλείται για κάθε επιμέρους τμήμα (γραμμές 7, 8). Τα $fingerprints$ που επιστρέφονται συνενώνονται στο $fingerprint$ του T.

6.6.3 Αυξητική (*incremental*) συμπίεση του *Evolution Graph*

Ο *batch* αλγόριθμος συμπίεσης της Εικόνας 6.9 απαιτεί ως είσοδο όλο τον *Evolution Graph*, πριν κατασκευάσει το $fingerprint$ του. Αυτή η προσέγγιση απαιτεί αρκετούς πόρους καθώς ο γράφος αυξάνεται συνεχώς. Έχουμε έτσι σχεδιάσει και μία *online* έκδοση του αλγορίθμου, τον $Incremental\ FINGERPRINT$, ο οποίο συμπιέζει τα $traces$ αυξητικά και δεν απαιτεί την εκ των προτέρων κατασκευή του *Evolution Graph*. Παρουσιάζουμε τον $Incremental\ FINGERPRINT$ στην Εικόνα 6.10.

$Incremental\ FINGERPRINT(FEG)$ Input: FEG // the fingerprint built so far
ζ // the most recent clustering, build at timepoint t_{i-1}
ξ // the current clustering, build at the current timepoint t_i
Output: FEG // the updated fingerprint

1. E_i=$MONIC(\zeta, \xi)$;
2. for each edge$e = (x, y) \in E_i$ do
3. if $e.extTrans \neq$ "survival" then
4. $FEG.addNode(y)$;
5. $FEG.addEdge(e)$;
6. else if $dist(x.label, \widehat{y}) \geq \tau$ then // $C2$ is violated
7. $FEG.addNode(y)$;
8. $FEG.addEdge(e)$;
9. else
10. $v = vcenter(x, y)$;
11. $FEG.replaceNode(x, v)$;
12. endif
13. end-for
14. return FEG;

Figure 6.10: Ο αλγόριθμος $Incremental\ FINGERPRINT$ για την *online* κατασκευή και συμπίεση του *Evolution Graph*

Ο $Incremental\ FINGERPRINT$ καλεί το $MONIC$ (γραμμή 1), το οποίο συγκρίνει την τρέχουσα συσταδοποίηση ξ (την t_i) με την πιο πρόσφατη συσταδοποίηση ζ (την t_{i-1}), εντοπιζει τις μεταβολές συστάδων και τις επιστρέφει ως ένα σύνολο ακμών E_i, με βάση την Ενότητα 6.4. Ο κόμβος - αρχή κάθε ακμής αντιστοιχεί σε έναν κόμβο που βρίσκεται ήδη στον FEG.

Για κάθε ακμή $e = (x, y)$, ο αλγόριθμος *Incremental FINGERPRINT* εξετάζει κατά πόσο η e αντιστοιχεί σε επιβίωση (γραμμή 3), δηλαδή, κατά πόσο η e αποτελεί τμήμα κάποιου *trace*. Αν όχι, ο FEG επεκτείνεται προσθέτοντας τη συστάδα y και την ακμή e (γραμμές 4, 5). Ωστόσο, δεν προσθέτουμε όλη τη συστάδα παρά μόνο την ετικέτα της (βλέπε Ενότητα 6.5.1.1).

Αν η $e = (x, y)$ δεν ανήκει σε κάποια *trace*, ο *Incremental FINGERPRINT* ελέγχει κατά πόσο οι ετικέτες των x και y είναι παρόμοιες, με βάση τη συνθήκη *(C2)* του Ορισμού 26 (γραμμή 6). Δεδομένου ότι η συστάδα x έχει ήδη προστεθεί στον FEG, προσπελαύνουμε την ετικέτα της κατευθείαν, ενώ η ετικέτα της συστάδας y, πρέπει να υπολογιστεί. Αν η συνθήκη *(C2)* δεν ικανοποιείται, ο FEG επεκτείνεται με την προσθήκη των y και e, όπως προηγουμένως. Αν τέλος, η $C2$ ικανοποιείται, τότε τα y και e δεν χρειάζεται να προστεθούν στον FEG. Αντί αυτού, τα x και y συνοψίζονται στο εικονικό τους κέντρο v (γραμμή 10) και ο κόμβος x αντικαθίσταται από το v (γραμμή 11). Αυτό σημαίνει πως όλες οι ακμές που έδειχναν στην x πρέπει να κατευθυνθούν στο v.

Ο *Incremental FINGERPRINT* δουλεύει αυξητικά (*incrementally*), λαμβάνοντας υπόψη κατά τη συμπίεση μόνο ζεύγη από γειτονικούς κόμβους, αντί για ολόκληρα *traces*, δηλαδή, δουλεύει τοπικά και όχι καθολικά. Ωστόσο, έχει το πλεονέκτημα ότι δεν απαιτεί την εκ των προτέρων κατασκευή του *Evolution Graph*.

6.7 Πειραματική μελέτη

Εφαρμόσαμε το $MONIC+$ σε ένα συνθετικό σύνολο δεδομένων (Ενότητα 6.7.1). Επίσης, εφαρμόσαμε το $MONIC$ σε μία πραγματική συλλογή από κείμενα, την Ενότητα H2.8 της ACM βιβλιοθήκης για το διάστημα από το 1997 μέχρι το 2004 (Ενότητα 6.7.2). Στόχος των πειραμάτων μας με το συνθετικό σύνολο δεδομένων είναι να δείξουμε τη δυναμικότητα του πλασίου $MONIC+$και να δείξουμε το είδος των μεταβολών που μπορεί να εντοπίσει, για το σκοπό αυτό χρησιμοποιήσαμε ένα $2 - D$ σύνολο δεδομένων το οποίο μπορεί εύκολα να οπτικοποιηθεί. Στόχος των πειραμάτων μας με το πραγματικό σύνολο δεδομένων είναι να κατανοήσουμε την εξέλιξη ενός πραγματικού πληθυσμού και να μελετήσουμε την επίδραση των διαφόρων παραμέτρων στη διαδικασία εντοπισμού μεταβολών.

Πειραματιστήκαμε επίσης με το πλαίσιο $FINGERPRINT$ σε τρία σύνολα δεδομένων και μελετήσαμε την απώλεια πληροφορίας και το κέρδος συμπαγότητας κατά τη διαδικασία συμπίεσης της ιστορίας της εξέλιξης του πληθυσμού (Ενότητα6.7.3).

6.7.1 Πειράματα στο $MONIC+$

Πειραματιστήκαμε με το $MONIC+$ σε ένα συνθετικό ρεύμα δεδομένων, στο οποίο εισάγαμε διάφορες μεταβολές (Ενότητα 6.7.1.1). Αναφέρουμε εδώ τα αποτελέσματα σε συστάδες τύπου B1 (Ενότητα 6.7.1.2) και τύπου A (Ενότητα 6.7.1.3).

6.7.1.1 Περιγραφή των δεδομένων

Χρησιμοποιούμε ένα γεννήτορα ο οποίος παίρνει σαν είσοδο το πλήθος των δεδομένων M, το πλήθος των συστάδων K, καθώς επίσης και το μέσο και την τυπική απόκλιση κάθε συστάδας. Τα δεδομένα γεννιούνται γύρω από ένα μέσο

και με βάση κάποια τυπική απόκλιση, ακολουθώντας μία *Gaussian* κατανομή. Σταθεροποιήσαμε την τυπική απόκλιση στο 5 και χρησιμοποιήσαμε ένα $2D$ χώρο μεγέθους 100×100 για τη δημιουργία. Το ρεύμα δεδομένων δημιουργήθηκε όπως περιγράφεται ακολούθως (βλέπε Εικόνα 6.3).

- t_1**:** Το σύνολο δεδομένων d_1 αποτελείται από σημεία γύρω παό τα $K_1 = 5$ κέντρα (20,20), (20, 80), (80, 20), (80, 80), (50, 50).

- t_2**:** Το σύνολο δεδομένων d_2 αποτελείται από 40 σημεία, ομοιόμορφα κατανεμημένα γύρω από τα κέντρα των τεσσάρων γωνιακών συστάδων του συνόλου δεδομένων d_1.

- t_3**:** Το σύνολο δεδομένων d_3 αποτελείται από 30 σημεία γύρω από το σημείο (50,40) και από 30 σημεία γύρω από το (50,60).

- $t_4, \ldots$**:** Τις t_4, t_5, t_6 προσθέσαμε 30 σημεία γύρω από τα t_4 :(20,50), t_5 :(20,30) και t_6 :(20,40).

Όσον αφορά στη γήρανση των δεδομένων, χρησιμοποιήσαμε μέγεθος παραθύρου $w = 2$. Συνεπώς, σε κάθε $t_i, i > 1$, το σύνολο δεδομένων είναι $D_i = d_i \cup d_{i-1}$.

6.7.1.2 Εντοπισμός μεταβολών για συστάδες τύπου B1

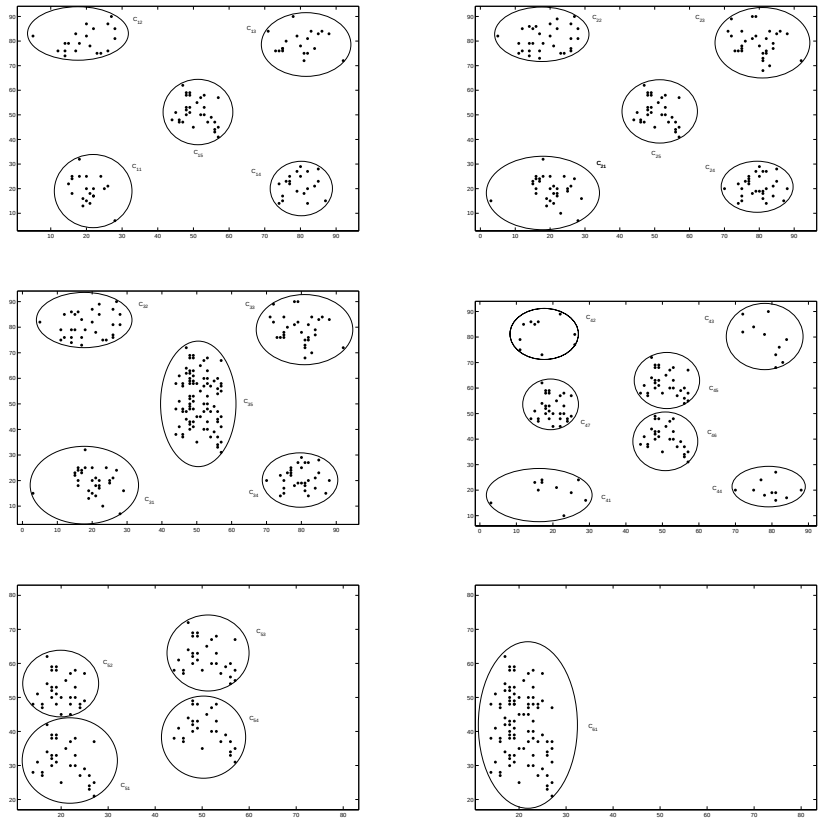
Δημιουργήσαμε τις συστάδες τύπου B1 πάνω στα σύνολα δεδομένων $D_1, \ldots, D_6$ και μελετήσαμε τη μετάβολή τους χρησιμοποιώντας τους δείκτες μεταβολών που παρουσιάστηκαν στην Ενότητα 6.3, για το πλαίσιο $MONIC$. Θέσαμε $\tau \equiv \tau_{match} = 0.5$, $\tau_{split} = 0.2$ και $\varepsilon = 0.003$.

Τα αποτελέσματα της συσταδοποίησης $\zeta_i, i = 1 \ldots 6$ παρουσιάζονται στην Εικόνα 6.11. Η εικόνα αυτή παρουσιάζει τις συστάδες σε κάθε χρονική στιγμή αλλά παρέχει λίγη πληροφορία σε σχέση με την επίδραση των νέων εγγραφών που προστίθενται στο σύνολο δεδομένω και της συνάρτησης γήρανσης. Στον Πίνακα 6.8, οι αλλαγές στον πληθυσμό ανταναλώνται στις μεταβολές που έχουν εντοπιστεί.

	Μεταβολές συστάδων				
t_2	$C_{11} \nearrow C_{21}$	$C_{12} \nearrow C_{22}$	$C_{13} \nearrow C_{23}$	$C_{14} \nearrow C_{24}$	$C_{15} \rightarrow C_{25}$
t_3	$C_{21} \rightarrow C_{31}$	$C_{22} \rightarrow C_{32}$	$C_{23} \rightarrow C_{33}$	$C_{24} \rightarrow C_{34}$	$C_{25} \nearrow C_{35}$
t_4	$C_{31} \rightarrow \odot$ $\odot \rightarrow C_{41}$	$C_{32} \rightarrow \odot$ $\odot \rightarrow C_{42}$	$C_{33} \rightarrow \odot$ $\odot \rightarrow C_{43}$	$C_{34} \rightarrow \odot$ $\odot \rightarrow C_{44}$	$C_{35} \overset{\subseteq}{\rightarrow} \{C_{45}, C_{46}\}$ $\odot \rightarrow C_{47}$
t_5	$C_{41} \rightarrow \odot$	$C_{42} \rightarrow \odot$	$C_{43} \rightarrow \odot$ $\odot \rightarrow C_{51}$	$C_{44} \rightarrow \odot$ $C_{46} \rightarrow C_{54}$	$C_{45} \rightarrow C_{53}$ $C_{47} \rightarrow C_{52}$
t_6	$C_{51} \overset{\subseteq}{\rightarrow} C_{61}$	$C_{52} \overset{\subseteq}{\rightarrow} C_{61}$	$C_{53} \rightarrow \odot$	$C_{54} \rightarrow \odot$	

Πίνακας 6.8: Μεταβολές συστάδων για συστάδες τύπου B1

Όπως μπορεί να δει κανείς, υπάρχουν τόσο εξωτερικές όσο και εσωτερικές μεταβολές. Συγκρίνοντας τα στοιχεία αυτού του πίνακα με την οπτικοποίηση της Εικόνα 6.11, μπορεί κανείς να δει πως το $MONIC$ σωστά αντιστοίχισε τις παλιές συστάδες στις καινούριες, εντοπίζοντας επιβιώσεις με ή χωρίς εσωτερικές μεταβολές, απορροφήσεις και διασπάσεις. Μερικές συστάδες υπέστησαν πολλαπλές



Σχήμα 6.11: Συστάδες τύπου B1 (μέσω του EM) τις χρονικές στιγμές t_1, t_2 (επάνω)· t_3, t_4 (μέση) και t_5, t_6 (κάτω)

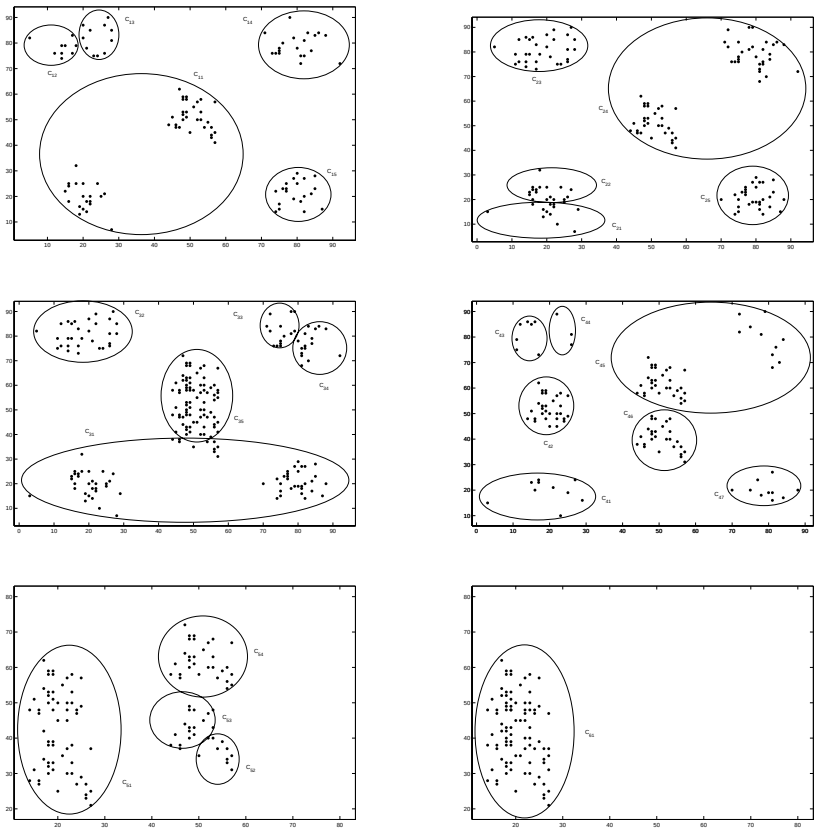
εσωτερικές μεταβολές , π.χ., η C_{12} διογκώθηκε ως προς το μέγεθός της και μετατοπίστηκε στη συστάδα C_{22}, η οποία επιπλέον, είναι πιο συμπαγής σε σχέση με την C_{12}. Υπάρχουν επίσης νέες συστάδες στις χρονικές στιγμές t_4 και t_5.

6.7.1.3 Εντοπισμός μεταβολών για συστάδες τύπου Α

Δημιουργήσαμε τις συστάδες τύπου Α πάνω στα σύνολα δεδομένων $D_1, \ldots, D_6$ μέσω του αλγορίθμου $K-means$ χρησιμοποιώντας για τιμή του K τη βέλτιστη τιμή συστάδων που ανακάλυψε ο αλγόριθμος EM.

Τα αποτελέσματα της συσταδοποίησης $\zeta_i, i = 1, \ldots, 6$ παρουσιάζονται στην Εικόνα 6.12. Τα αποτελέσματα αυτά διαφέρουν σε σχέση με τα αποτελέσματα του EM, συνεπώς υπάρχουν και διαφορετικές μεταβολές συστάδων. Χρησιμοποιήσαμε τους δείκτες μεταβολών του Πίνακα 6.5, θέτοντας $\tau = 0.5$ και $\tau_{split} = 0.2$. Για τη μεταβολή μεγέθους, χρησιμοποιήσαμε το δείκτη B1 του Πίνακα 6.3 με $\varepsilon = 0.003$. Για τις υπόλοιπες εσωτερικές μεταβολές, χρησιμοποιήσαμε τους δείκτες του Πίνακα 6.6 με $\tau_{location} = 0.1$ (για την μεταβολή τοποθεσίας) και $\varepsilon = 0.001$ (για την μεταβολή συμπαγότητας).

Οι μεταβολές συστάδων που εντοπίστηκαν από το $MONIC+$ φαίνονται σ-



Σχήμα 6.12: Συστάδες τύπου Α (μέσω του $K-means$) τις χρονικές στιγμές t_1, t_2· t_3, t_4 και t_5, t_6

τον Πίνακα 6.9 και αποκαλύπτουν ότι οι περισσότερες συστάδες είναι με ασταθείς και αντιμετωπίζουν εσωτερικές αλλαγές όλων των ειδών ή εξαφανίζονται. Ακόμα και χωρίς τη δυνατότητα οπτικοποίησης των αποτελεσμάτων (που μπορεί να είναι δύσκολη σε πολυδιάστατους χώρους, οι μεταβολές αυτές δείχνουν την αστάθεια των συστάδων και την ανάγκη επισταμένης μελέτης των επιμέρους συστάδων.

Συγκρίνοντας τις μεταβολές συστάδων που προέκυψαν μέσω του $K-means$ (Εικόνα 6.12) και του EM (Εικόνα 6.11), μπορεί κανείς να παρατηρήσει πως οι συστάδες του $K-means$ αντιμετωπίζουν πιο πολλές μεταβολές σε σχέση με αυτές του EM, γεγονός που οφείλεται στο ότι ο $K-means$ οδηγεί σε λιγότερο ευσταθείς συστάδες.

6.7.2 Πειράματα στο $MONIC$

Πειραματιστήκαμε με το $MONIC$ στη βιβλιοθήκη της ACM και συγκεκριμένα στην κατηγορία $H2.8$ (database applications) με στόχο να αποκτήσουμε κατανόηση σε σχέση με την εξέλιξη του πληθυσμού και να δούμε πως επηρεάζεται το πλαίσιο από τις διάφορες παραμέτρους.

Περιγράφουμε πρώτα το σύνολο δεδομένων (Ενότητα 6.7.2.1) και εν συνεχεία

	Μεταβολές συστάδων				
t_2	$C_{11} \rightarrow \odot$	$C_{12} \stackrel{\subseteq}{\rightarrow} C_{23}$	$C_{13} \stackrel{\subseteq}{\rightarrow} C_{23}$ $\odot \rightarrow C_{21}$	$C_{14} \rightarrow \odot$ $\odot \rightarrow C_{22}$	$C_{15} \cdots \stackrel{\bullet}{\rightarrow} \nearrow C_{25}$ $\odot \rightarrow C_{24}$
t_3	$C_{21} \rightarrow \odot$	$C_{22} \rightarrow \odot$ $\odot \rightarrow C_{31}$	$C_{23} \rightarrow C_{32}$ $\odot \rightarrow C_{33}$	$C_{24} \rightarrow \odot$ $\odot \rightarrow C_{34}$	$C_{25} \rightarrow \odot$ $\odot \rightarrow C_{35}$
t_4	$\odot \rightarrow C_{41}$	$\odot \rightarrow C_{47}$	$C_{33} \rightarrow \odot$	$C_{34} \rightarrow \odot$ $\odot \rightarrow C_{42}$	$C_{35} \cdots \stackrel{\star}{\rightarrow} \searrow C_{45}$ $C_{32} \stackrel{\subseteq}{\rightarrow} \{C_{43}, C_{44}\}$ $C_{31} \cdots \stackrel{\bullet}{\rightarrow} \searrow C_{46}$
t_5	$C_{41} \rightarrow \odot$	$C_{47} \rightarrow \odot$	$C_{43} \rightarrow \odot$	$C_{44} \rightarrow \odot$	$C_{45} \cdots \stackrel{\bullet}{\rightarrow} \searrow C_{54}$ $C_{46} \stackrel{\subseteq}{\rightarrow} \{C_{52}, C_{53}\}$ $C_{42} \cdots \stackrel{\star}{\rightarrow} \nearrow C_{51}$
t_6		$C_{52} \rightarrow \odot$	$C_{53} \rightarrow \odot$	$C_{54} \rightarrow \odot$	$C_{51} \stackrel{\bullet}{\rightarrow} \nearrow C_{61}$

Πίνακας 6.9: Μεταβολές συστάδων για συστάδες τύπου Α

παρουσιάζουμε τα αποτελέσματα (Ενότητα 6.7.2.2).

6.7.2.1 Περιγραφή της Ενότητας $H2.8$ της ACM βιβλιοθήκης

Η ενότητα $H2.8$ (Database applications") περιέχει δημοσιεύσεις σε (1) εξόρυξη γνώσης, (2) χωρικές βάσεις δεδομένων, (3) βάσεις εικόνων, (4) στατιστικές βάσεις δεδομένων, (5) επιστημονικές βάσεις δεδομένων, κατηγοριοποιημένες στις αντίστοιχες κλάσεις. Περιέχει επίσης (6) μη-κατηγοριοποιημένες δημοσιεύσεις, δηλαδή, αυτές που ανήκουν στην βασική κλάση "Database applications".

Για το διάστημα 1997 μέχρι 2004, επιλέξαμε άρθρα που οι κλάσεις τους ανήκαν σε κάποια από αυτές τις 6 κλάσεις της $H2.8$. Να αναφέρουμε εδώ πως η συλλογή δεν έχει ισορροπία: η κλάση (1) είναι μεγαλύτερη από όλες τις άλλες μαζί και μεγαλώνει πιο γρήγορα από αυτές. Για κάθε άρθρο, θεωρήσαμε τον τίτλο του και μία λίστα από λέξεις κλειδιά. Για τη συσταδοποίηση χρησιμοποιήσαμε το σχήμα $TFIDF$ [;], τις 30 πιο συχνές λέξεις ως χώρο γνωρισμάτων και τον αλγόριθμο bisecting K-means για $K = 10$. Η συσταδοποίηση έγινε μέσω του DIAsDEM [19]. Για την γήρανση των δεδομένων, χρησιμοποιήσαμε μέγεθος παραθύρου 2.

Οι μεταβολές που εντόπισε το πλαίσιο $MONIC$ παρουσιάζονται στην επόμενη ενότητα.

6.7.2.2 Μεταβολές συστάδων και επίδραση των κατωφλίων

Μεταβάλαμε το κατώφλι τ_{match} από 0.45 μέχρι 0.7 με βήμα 0.05 και βρήκαμε το πλήθος των συστάδων που υπέστησαν εσωτερικές ή εξωτερικές αλλαγές. Για τ_{match} μεγαλύτερο του 0.7, δεν υπήρχαν συστάδες που να επέζησαν. Θέσαμε επίσης το $\tau_{split} = 0.1$. Τα αποτελέσματα φαίνονται στην Εικόνα Φιγ. 6.13.

Στην Εικόνα 6.13(α) βλέπουμε ότι καθώς το τ_{match} αυξάνεται και γίνεται πιο περιοριστικό το πλήθος των συστάδων που επιζούν μειώνεται. Το πλήθος των διασπάσεων και εξαφανίσεων στην Εικόνα 6.13 και(γ) αντίστοιχα, αυξάνεται. Την ίδια στιγμή όλες οι συστάδες που επιζούν παρουσιάζουν εσωτερικές μεταβολές

μεγέθους. Επιπλέον, δεν εντοπίστηκαν τυχόν απορροφήσεις συστάδων, συνεπώς ο ρυθμό $passforward$ ισούται με το ρυθμό επιβίωσης για όλες τις συσταδοποιήσεις.

Συγκρίνοντας τις τιμές σε κάθε χρονική στιγμή, η Εικόνα 6.13 (α), (β), (γ), αποκαλύπτει πως οι συστάδες στις αρχικές συσταδοποιήσεις εξαφανίζονται και αντικαθίστανται από νέες συστάδες (πιο πολλές εξαφανίσεις από διασπάσεις), ενώ αυτή η τάση αντιστρέφεται στις τελευταίες συσταδοποιήσεις: Οι συστάδες στις πιο πρόσφατες συσταδοποιήσεις κυρίως διασπώνται παρά εξαφανίζονται. Αυτό μπορεί να οφείλεται στον αυξανόμενο όγκο της συλλογής: Το πλήθος των άρθρων που εισάγεται σε κάθε χρονικής στιγμή αυξάνεται ραγδαία στις τελευταίες χρονικές στιγμές, με αποτέλεσμα μη σταθερές συστάδες να διασπώνται από τον αλγόριθμο συσταδοποίησης σε μεγάλα κομμάτια αντί να διαλύονται και να ξαναδημιουργούνται.

Για να ελέγξουμε αυτή την υπόθεση, αναλύσαμε την επίδραση του τ_{split} στο πλήθος των διασπάσεων και εξαφανίσεων όπως φαίνεται στην Εικόνα 6.14. Μεταβάλαμε το τ_{split} από 0.1 έως 0.35 με βήμα 0.05, σταθεροποιώντας το τματςη $= 0.5$. Όπως αναμενόταν, μεγάλες τιμές του τ_{split} οδηγούν σε μεγάλο αριθμό εξαφανίσεων. Ωστόσο, το πλήθος των διασπάσεων στις τελευταίες χρονικές στιγμές δείχνουν ότι μία διάσπαση είναι δυνατή αν η τιμή του τ_{split} είναι μικρή. Πράγματι, οι συστάδες δεν διασπώνται σε μεγάλα κομμάτια, αλλά αποσυντίθενται και ξαναδημιουργούνται. Για παράδειγμα, η επικρατούσα κλάση (" Εξόρυξη Γνώσης ") αυξάνεται σημαν- τικά τις τελευταίες χρονικές στιγμές αλλά δεν είναι αρκετά ομοιογενής ώστε να δημιουργήσει συστάδες με μεγάλο χρόνο ζωής.

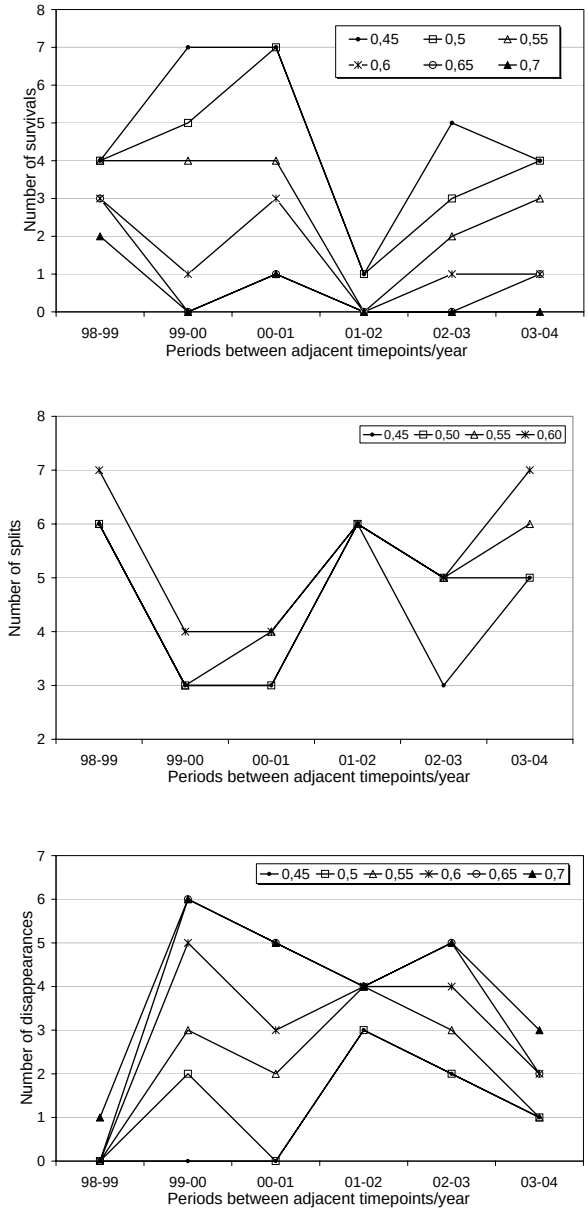
6.7.2.3 Χρόνος ζωής συστάδων και συσταδοποιήσεων

Μελετάμε στη συνέχεια το χρόνο ζωής συστάδων και συσταδοποιήσεων στην πορεία του χρόνου, τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα. Για λόγους απλότητας δείχνουμε τον ακριβή αριθμό συστάδων. Είναι εμφανές ότι υπάρχουν στιγμές με υψηλό ή χαμηλό ρυθμό $passforward$ ανεξάρτητα από την τιμή του τ_{match}: Η χαμηλή τιμή $passforward$ το 2002 δείχνει μία δραστική αλλαγή του πληθυσμού μεταξύ των ετών 2001 και 2002 (window size = 2), που έπεται μιας σταθερής περιόδου δύο ετών (οι συσταδοποιήσεις των 2000 και 2001 έχουν σχετικά υψηλούς ρυθμούς $passforward$).

τ	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*
0.45	4	7	7	1	5	4
0.50	4	5	7	1	3	4
0.55	3	3	3	0	2	3
0.60	3	2	3	0	1	1
0.65	3	0	1	0	0	1
0.70	2	0	1	0	0	0

Πίνακας 6.10: Οι ρυθμοί $passforwardratios$ για διάφορες τιμές του τ_{match}

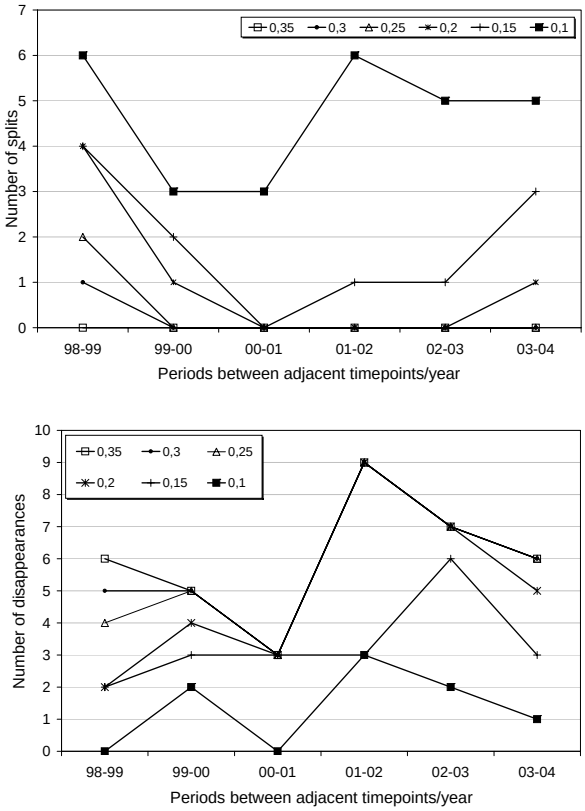
Οι εξαιρετικά χαμηλές τιμές του ρυθμού $passforward$ καθρεπτίζονται και στους χρόνους ζωής των επιμέρους συστάδων. Μελετήσαμε το χρόνο ζωής των συστάδων με βάση τον Ορισμό 18. Για το σκοπό αυτό, θέσαμε $\tau_{match} = 0.5$ και $\tau_{split} = 0.1$ και υπολογίσαμε το χρόνο ζωής με εσωτερικές μεταβολές για τις συστάδες, $lifetimeI$: Δεδομένου ότι όλες οι συστάδες που επιβίωσαν έχουν υποστεί εσωτερικές αλλαγές, ο αυστηρός χρόνος ζωής είναι 1 για όλες τις συστάδες.



Σχήμα 6.13: Μεταβολές συστάδων για διαφορετικές τιμές του τ_{match}: (α) επιβιώσεις, (β) διασπάσεις και (γ) εξαφανίσεις

Επιπλεόν, δεδομένου ότι δεν έχουν συμβεί απορροφήσεις, ο χρόνος ζωής με απορροφήσεις είναι επίσης με το $lifetimeI$ για όλες τις συστάδες. Τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα. Η δεύτερη στήλη αυτού του πίνακα δείχνει το χρόνο ζωής όλων των συστάδων σε κάθε συσταδοποίηση.

Η τρίτη στήλη στον Πίνακα 6.11 είναι ο χρόνος ζωής των συσταδοποιήσεων με



Σχήμα 6.14: Μεταβολές συστάδων για διαφορετικές τιμές του τ_{split}: (α) διασπάσεις και (β) εξαφανίσεις

Χρονική στιγμή	*Χρόνοι ζωής συστάδων*	*Χρόνος ζωής L*
1998	{4,1,4,1,4,1,2,1,1,1}	1
1999	{4,1,4,1,1,4,1,2,1,3}	1
2000	{4,1,4,1,3,2,4,3,2,1}	2
2001	{4,1,4,2,4,2,3,2,1,1}	2
2002	{1,1,3,1,1,1,1,1,1,1}	1
2003	{3,1,1,1,1,1,1,2,3,1}	1
2004	{3,3,1,2,1,1,2,1,1,1}	1

Πίνακας 6.11: Χρόνος ζωής συσταδοποιήσεων

βάση τον Ορισμό 19. Είναι προφανές ότι όλες οι χρονικές στιγμές χαρακτηρίζονται από βραχυπρόθεσμες συσταδοποιήσεις, παρόλο που μερικές από αυτές περιέχουν μερικές ευσταθείς συστάδες.

6.7.2.4 Συστάδες εναντίον κλάσεων στην βιβλιοθήκη *ACM*

Μέχρι τώρα μελετήσαμε τις αλλαγές στην $H2.8$ χωρίς να λάβουμε υπόψη τις πραγματικές κλάσεις των άρθων. Στην ενότητα αυτή, αντιπαραθέτουμε τις μεταβολές και τους χρόνους ζωής των συστάδων που βρήκαμε μέσω του $MONIC$ με τις πραγματικές αλλαγές που παρατηρούνται στην $H2.8$. Για το σκοπό αυτό, αναθέσαμε μία ετικέτα σε κάθε συστάδα (η ετικέτα αυτή αποτελείται από τις δύο πιο συχνές λέξεις κλειδιά που εμφανίζονται στη συστάδα) και αντιστοιχίσαμε αυτές τις ετικέτες με τις πραγματικές ετικέτες/ θέματα των κλάσεων της ΑΜ. Για τον εντοπισμό των μεταβολών, θέσαμε $\tau = 0.5$ και $\tau_{split} = 0.1$.

Τα αποτελέσματα είναι τα ακόλουθα:

- Υπάρχει πάντα μία συστάδα χωρίς ετικέτα την οποία την αποκαλούμε *cluster* 0. Ο αλγόριθμος τοποθετεί σε αυτή εγγραφές που δεν μπορούν να τοποθετηθούν σε άλλες συστάδες. Από τη φύση της η συστάδα αυτή έχει μεγάλο χρόνο ζωής, επιβιώνει για 4 χρονικές στιγμές. Και αυτή όμως, επηρεάζεται από την μετατόπιση του πληθυσμού το 2002· τη συγκεκριμένη χρονική στιγμή διαλύεται και επαναδημιουργείται.

- Κάθε συσταδοποίηση περιέχει 2 ή 3 συστάδες σε εξόρυξη γνώσης, την κυρίαρχη κλάση του πληθυσμού. Στις πρώτες 4 χρονικές στιγμές, υπήρχε μία συστάδα σε association rules, η οποία διασπάστηκε το 2002 σε μία μικρότερη συστάδα με την ίδια ετικέτα και μία συστάδα με θόρυβο (διαφορετική από τη συστάδα 0):

$$C_{1998_4} \nearrow C_{1999_9} \nearrow C_{2000_6} \nearrow C_{2001_4} \overset{\subseteq}{\rightarrow} \{C_{2002_7}, C_{2002_9}\}$$

  όπου το C_{y_w} συμβολίζει το αναγνωριστικό της συστάδας "association rules" στο έτος y, $w = 1 \ldots 9$. (Τα αναγνωριστικά των συστάδων δημιουργούνται από τον αλγόριθμο συσταδοποίησης σε κάθε χρονική στιγμή. Δεν υποδηλώνουν μεταβολές). Η μικρή συστάδα C_{2002_7} εξαφανίζεται το 2003 ($C_{2002_7} \rightarrow \odot$). Μία από τις προκύπτουσες συστάδες το 2004 ($\odot \rightarrow C_{2004_3}$) έχει πάλι την ετικέτα association rules.

- Οι άλλες συστάδες σε εξόρυξη γνώσης εχουν πιο γενικές ετικέτες όπως knowledge discovery ή data mining. Ο χρόνος ζωής τους δεν υπερβαίνει τις 3 χρονικές στιγμές, κατά τη διάρκεια των οποίων υπόκεινται σε εσωτερικές μεταβολές.

- Το πλήθος των συστάδων για το οποίο υπάρχουν ετικέτες είναι μεγαλύτερο στις αρχικές συσταδοποιήσεις και μειώνεται στις πιο πρόσφατες συσταδοποιήσεις. Οι ετικέτες στις τελευταίες συσταδοποίησεις είναι πιο σύντομες και με λιγότερη πληροφορία($<< model >>$,$<< data >>$). Συστάδες που μπορούν να ανατεθούν σε κλάσεις άλλες από την Data Mining εμφανίζονται μόνο στις αρχικές χρονικές στιγμές.

Συνεπώς, το $MONIC$ βρήκε μία αξιοσημείωτη μετατόπιση του πληθυσμού της H2.8 μεταξύ του 2001 και 2002 η οποία σημαδεύτηκε από έναν αυξημένο αριθμό διασπάσεων και εξαφανίσεων. Η ιστορία της H2.8 περιέχει ένα τουλάχιστον γεγονός που μπορεί να δικαιολογεί αυτή τη μετατόπιση: Αρχίζοντας από το KDD 2001, οι εργασίες του συνεδρίου KDD και κάποιων *workshops* του άρχισαν να ανεβαίνουν στην ψηφιακή βιβλιοθήκη της ACM, εμπλουτίζοντάς την με πολλά άρθρα σχετικά με εξόρυξη γνώσης.

6.7.3 Πειράματα στο $FINGERPRINT$

Πρώτα περιγράφουμε τα σύνολα δεδομένων που χρησιμοποιήσαμε στα πειράματα και παρέχουμε κάποια ενδεικτικά παραδείγματα από *traces* και τα αντίστοιχά τους $fingerprints$. Στη συνέχεια, συζητάμε τα αποτελέσματα για κάθε σύνολο δεδομένων.

6.7.3.1 Περιγραφή των δεδομένων

Πειραματιστήκαμε με δύο αριθμητικά σύνολα δεδομένανω, το *Network Intrusion dataset* και το *Charitable Donation dataset*, που χρησιμοποιήθηκαν και στα πειράματα του[2], και με το *ACM H2.8 dataset* που αναφέραμε και στα πειράματα του $MONIC$ (βλέπε Ενότητα 6.7.2.1). Το πρώτο σύνολο δεδομένων είναι εξαιρετικά δυναμικό, το δεύτερο σχετικά στατικό, ενώ το τρίτο εξελίσσεται με μη συμμετρικό τρόπο, καθώς μία κλάση (η "Data Mining") μεγαλώνει πιο γρήγορα από τις άλλες.

Το *Network Intrusion dataset* (KDD Cup'99) περιέχει *logs* από TCP συνδέσεις σε ένα δίκτυο LAN κατά τη διάρκεια 2 εβδομάδων (424,021 εγγραφές). Κάθε εγγραφή αντιστοιχεί σε μία κανονική σύνδεση ή σε μία επίθεση. Οι επιθέσεις ανήκουν σε μία από τις τέσσερις μεγάλες κατηγορίες: DOS (δηλαδή, $denial-of-service$), $R2L$ (δηλαδή, unauthorized access from a remote machine), $U2R$ (δηλαδή, unauthorized access to local superuser privileges), και $PROBING$ (δηλαδή, surveillance and other probing). Συνεπώς, θέσαμε το πλήθος των συστάδων στις 5, λαμβάνοντας υπόψη και την κλάση των κανονικών συνδέσεων. Για τη συσταδοποίηση, χρησιμοποιήσαμε και τα 34 συνεχή γνωρίσματα και απομακρύναμε ένα ακραίο σημείο, όπως στην[2]. Μετατρέψαμε το σύνολο δεδομένων σε ρεύμα δεδομένων ταξινομώντας τα δεδομένα με βάση το πότε εισήχθηκαν. Θεωρήσαμε μία ομοιόμορφη ροή με ταχύτητα 2000 στιγμιότυπα ανά χρονική στιγμή. Για τη γήρανση, θεωρήσαμε μέγεθος παραθύρου ίσο με 2.

Το *Charitable Donation dataset* (KDD Cup'98) περιέχει πληροφορίες (95,412 εγγραφές) σχετικά με ανθρώπους που συμμετείχαν σε φιλανθρωπίες. Η συσταδοποίηση αναγνωρίζει ομάδες ανθρώπων με παρόμοια φιλανθρωπική συμπεριφορά. Παρόμοια με την [21], χρησιμοποιήσαμε 56 γνωρίσματα από τα 481 και θέσαμε το πλήθος των συστάδων στις 10. Όπως και στο προηγούμενο σύνολο δεδομένων, χρησιμοποιήσαμε τη σειρά εισόδου των δεδομένων για τη διάταξη του ρεύματος δεδομένων και υποθέσαμε μία ομοιόμορφη ροη δεδομένων με 200 στιγμιότυπα σε κάθε χρονική περίοδο. Για το παράθυρο, θεωρήσαμε μέγεθος παραθύρου 2.

Το *ACM H2.8 dataset* το έχουμε ήδη περιγράψει στην Ενότητα 6.7.2.1, χρησιμοποιούμε και εδώ τις ίδιες ρυθμίσεις.

6.7.3.2 Ενδεικτικά παραδείγματα *traces* και $fingerprints$

Για να επιδείξουμε τη συμπεροράτης συμπίεσης στα πλαίσια του $FINGERPRINT$, παραθέτουμε κάποια παραδείγματα από *traces* από το σύνολο δεδομένων *ACM H2.8 dataset* και τα αντίστοιχα $fingerprint$ τους, όπως αυτά δημιουργήθηκαν από τους αλγορίθμους μας.

Το 1998, βλέπουμε μία συστάδα με ετικέτα "information systems". Το *trace* της συστάδας είναι το ακόλουθο:

$$trace(c_{1998_2}) = \prec c_{1998_2} c_{1999_6} c_{2000_3} \succ$$

όπου ο συμβολισμός c_{y_i} αναφέρεται στην i συστάδα του έτους y, για $i = 1 \ldots 9$. Τα *centroids* των συστάδων περιέχουν τους όρους "$information$" και "$system$"

με τις ακόλουθες συχνότητες:

$$\widehat{c_{1998_2}} =< information(0.96), system(0.61) >$$
$$\widehat{c_{1999_6}} =< information(0.88), system(0.74) >$$
$$\widehat{c_{2000_3}} =< information(0.76), system(0.78) >$$

Και οι δύο αλγόριθμοι συμπίεσης συμπίεσαν αυτό το *trace* σε ένα μόνο εικονικό κέντρο. Ο *batch* αλγόριθμος δημιούργησε αυτόν τον εικονικό κόμβο v σε ένα βήμα $\widehat{v} =< information(0.87), system(0.71) >$, ενώ ο *online* αλγόριθμος πρώτα συμπίεσε τις c_{1998_2} και c_{1999_6} στο εικονικό κέντρο $\widehat{v_0} =< information(0.92), system(0.68) >$, και στη συνέχεια συμπίεσε τα v_0 και c_{2000_3} στο $\widehat{v'} =< information(0.84), system(0.73) >$.

Μία άλλη συστάδα που αναδύθηκε το 1998 έχει ετικέτα $< analysis(1.0) >$. Το 1999, διασπάστηκε σε δύο συστάδες, η μία με ετικέτα $< mining(1.0), datum(0.74) >$ και η άλλη χωρίς ετικέτα (garbage cluster). Η πρώτη επιβίωσε για δύο περιόδους και το *trace* της είναι:

$$\prec c_{1999_8} c_{2000_4} c_{2001_6} \succ$$

Η πληροφορία που παίρνουμε από το γράφο χωρίς συμπίεση είναι η ακόλουθη:

$$c_{1998_9} \xrightarrow{\subseteq} \{c_{1999_4}, \prec c_{1999_8} c_{2000_4} c_{2001_6} \succ\}$$

Το *fingerprint* που προκύπτει μετά τη συμπίεση είναι το ακόλουθο:

$$c_{1998_9} \xrightarrow{\subseteq} \{c_{1999_4}, v\}$$

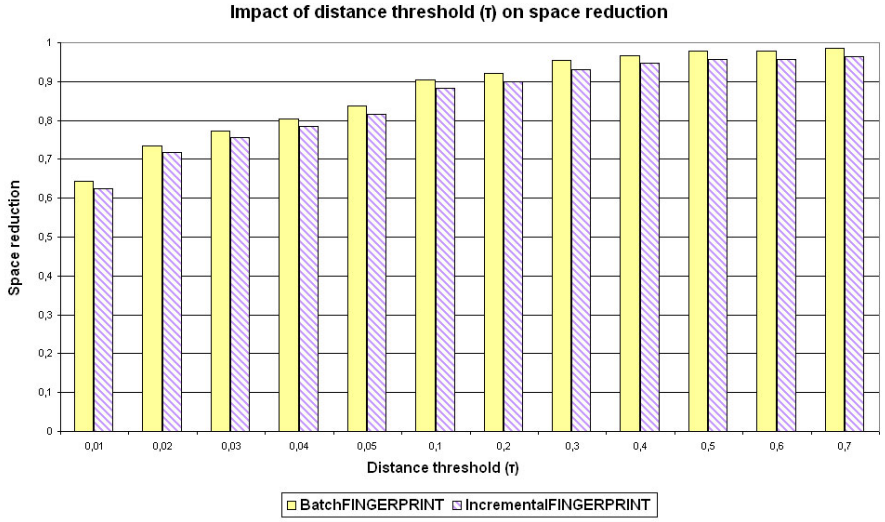
6.7.3.3 Κέρδος συμπαγότητας και απώλεια πληροφορίας

Στις Εικόνες 6.15, 6.16 και 6.17, δείχνουμε το κέρδος συμπαγότητας που επιτυγχάνουν οι αλγόριθμοι *Batch FINGERPRINT* και *Incremental FINGERPRINT* για καθένα από τα τρία σύνολα δεδομένωνμ θεωρώντας διαφορετικές τιμές για το άνω όριο δ στην απόσταση των *centroids*. Οι τιμές για τις διάφορες διαστάσεις των *centroids* βρίσκονται μεταξύ 0 και 1, έτσι μεταβάλαμε το δ σε αυτό το εύρος επίσης.

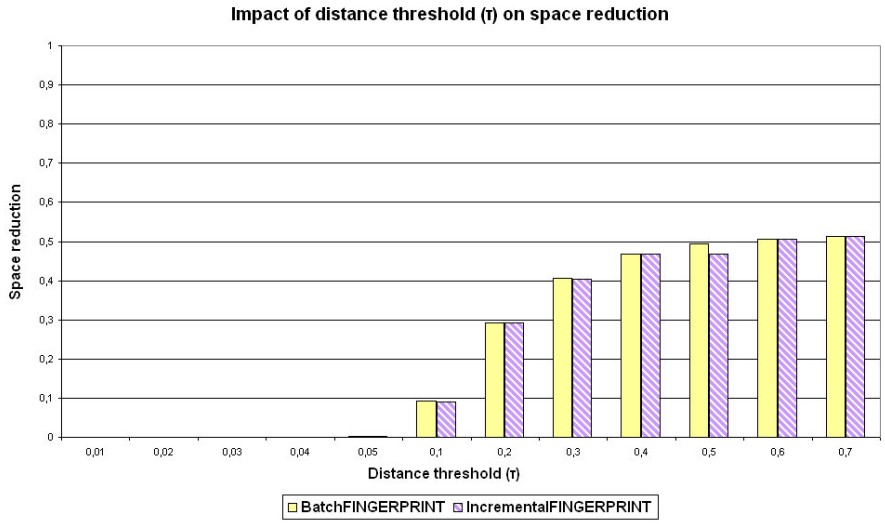
Μπορούμε να δούμε στις εικόνες αυτές πως οι δύο αλγόριθμοι επιτυγχάνουν παρόμοιο κέρδος συμπαγότητας, αν και ο αλγόριθμος *Incremental FINGERPRINT* παρουσιάζει ελαφρώς χαμηλότερες τιμές για τις περισσότερες τιμές του δ στο σύνολο δεδομένων *Network Intrusion dataset*. Προφανώς, καθώς το όριο απόστασης δ αυξάνεται, το κέρδος συμπαγότητας αυξάνεται επίσης (τα *centroids* συνενώνονται πιο εύκολα).

Το συνολικό κέρδος πληροφορίας για κάθε σύνολο δεδομένων εξαρτάται προφανώς από το πλήθος των επιβιώσεων: από το σύνολο των 1,195 συστάδων/κόμβων του Network Intrusion dataset, οι 400 κόμβοι συμμετέχουν σε *traces* και το κέρδος συμπαγότητας που επιτυγχάνουν και οι δύο αλγόριθμοι σε σχέση με το συνολικό μέγεθος του *Evolution Graph* βρίσκεται στο εύρος τιμών από 21% μέχρι 33%. Ο *Evolution Graph* του Charitable Donation dataset περιέχει 4,770 συστάδες, εκ των οποίων οι 614 συμμετέχουν σε *traces*. Το κέρδος συμπαγότητας πάνω σε όλο το γράφο δεν ήταν παραπάνω από 7%. Για το ACM H.2.8 dataset, οι 24 από τους 70 συνολικά κόμβους συμμετείχαν σε *traces*, έτσι το κέρδος συμπαγότητας πάνω σε όλο το γράφο ήταν μεταξύ του 9% και του 33%.

Στις Εικόνες 6.18, 6.19 και 6.20, παρουσιάζουμε την απώλεια πληροφορίας που δημιουργείται στα τρία σύνολα δεδομένων λόγω της συμπίεσης τόσο μέσω του

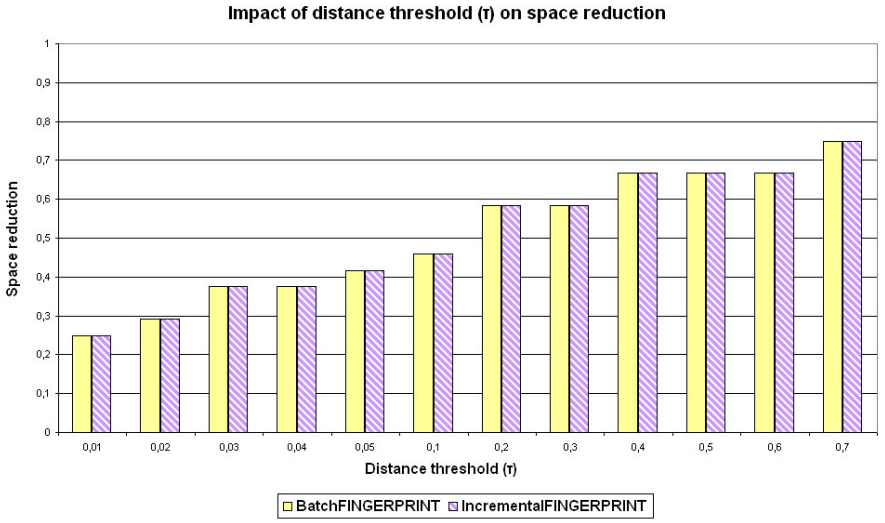


Σχήμα 6.15: Network Intrusion dataset: Επίδραση του κατωφλίου δ στο κέρδος συμπαγότητας



Σχήμα 6.16: Charitable Donation dataset: Επίδραση του κατωφλίου δ στο κέρδος συμπαγότητας

Batch FINGERPRINT όσο και μέσω του *Incremental FINGERPRINT* . Για το *Charitable Donation dataset* και το *ACM H2.8 dataset*, η απώλεια πληροφορίας που δημιουργεί ο *Incremental FINGERPRINT* είναι ελαφρώς υψηλότερη από την απώλεια πληροφορίας που δημιουργεί ο *Batch FINGERPRINT*

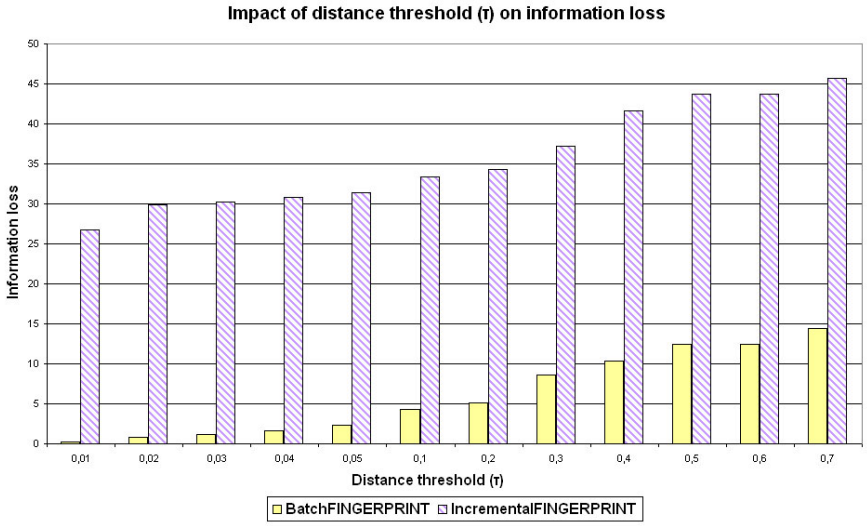


Σχήμα 6.17: ACM H.2.8 dataset: Επίδραση του κατωφλίου δ στο κέρδος συμπαγότητας

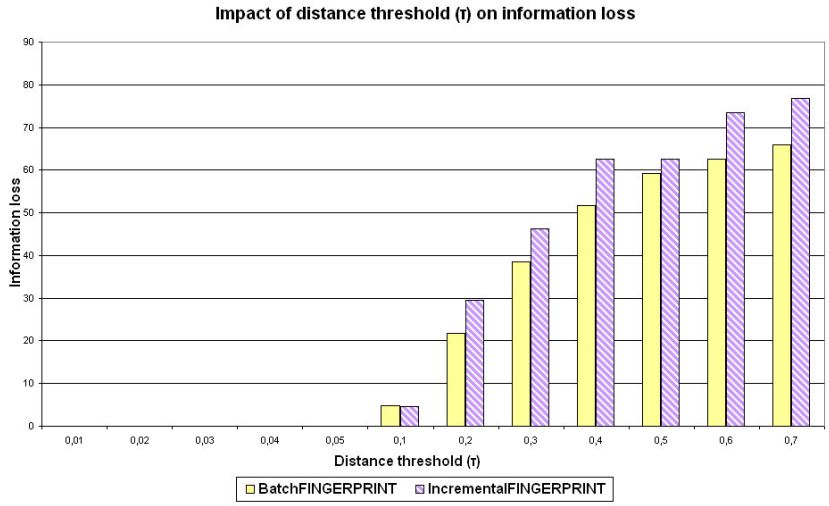
, αλλά ακολουθούν παρόμοια συμπεριφορά για τις διάφορες τιμές του κατωφλίου δ. Για το *Network Intrusion dataset*, η απόδοση διαφέρει δραστικά: Ενώ ο *Batch FINGERPRINT* καταφέρνει να επιτύχει μικρή απώλεια πληροφορίας (μικρότερη από ότι στα υπόλοιπα σύνολα δεδομένων), ο *Incremental FINGERPRINT* έχει πολύ χαμηλή απόδοση.

Συγκρίνοντας το κέρδος συμπαγότητας και την απώλεια πληροφορίας για τα σύνολα δεδομένων Charitable Donation dataset και *ACM H2.8*, μπορούμε να πούμε πως και οι δύο αλγόριθμοι συνοψίζουν τα *traces* με παρόμοιο τρόπο, αν και ο *Batch FINGERPRINT* αλγόριθμος καταφέρνει να συνοψίσει ελαφρώς περισσότερους κόμβους (υψηλότερο κέρδος συμπαγότητας) και πιο όμοιους μεταξύ τους (χαμηλότερη απώλεια πληροφορίας). Αυτό φαίνεται επίσης και στις Εικόνα 6.22 και Εικόνα 6.23, όπου δείχνουμε μαζί το κέρδος συμπαγότητας και την απώλεια πληροφορίας: Η συμπεριφορά των δύο αλγορίθμων είναι σχεδόν ίδια.

Η διαφορά στην απόδοση των δύο αλγορίθμων φαίνεται στην Εικόνα 6.21. Μία πιθανή εξήγηση για την χαμηλή απόδοση του *Incremental FINGERPRINT* θα μπορούσε να είναι η μεγάλη δυναμικότητα του συγκεκριμένου συνόλου δεδομένων (δηλαδή, Network Intrusion dataset): Το πλήθος των επιβιώσεων είναι σχετικά χαμηλό και είναι πιθανόν οι συστάδες που επιβιώνουν να είναι ασταθείς και αρκετά διαφορετικές μεταξύ τους. Συνεπώς, ο *Incremental FINGERPRINT* παρήγαγε εικονικά κέντρα που δεν ήταν αρκετά κοντά στα αρχικά ζεύγη των *centroids* των συστάδων, ενώ ο *Batch FINGERPRINT* κατάφερε να κατασκευάσει καλύτερα εικονικά κέντρα μεταξύ πολλών γειτονικών συστάδων. Περισσότερα πειράματα ωστόσο θα πρέπει να γίνουν εδώ. Συγκεκριμένα, θα πρέπει να εξεταστεί η ποιότητα των αρχικών συστάδων και πως αυτή σχετίζεται με το κέρδος συμπαγότητας και την απώλεια πληροφορίας που προκαλείται κατά την συμπίεση.



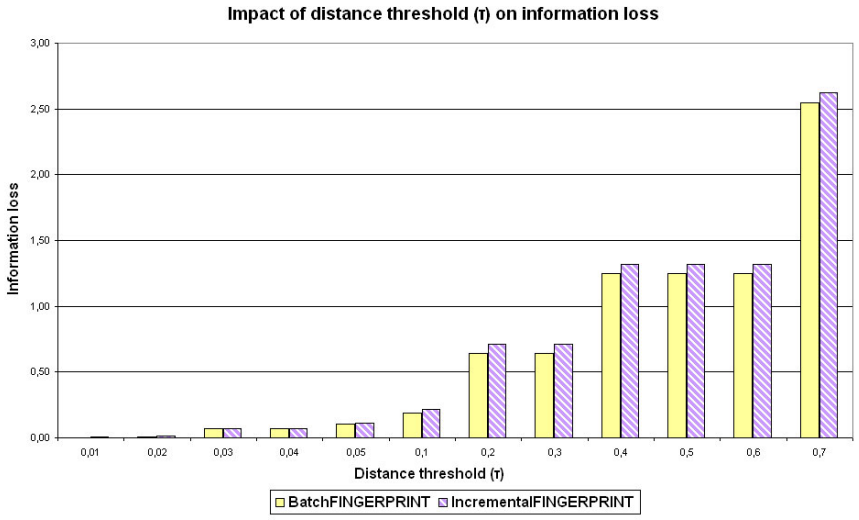
Σχήμα 6.18: Network Intrusion dataset: Επίδραση του κατωφίου δ στην απώλεια πληροφορίας



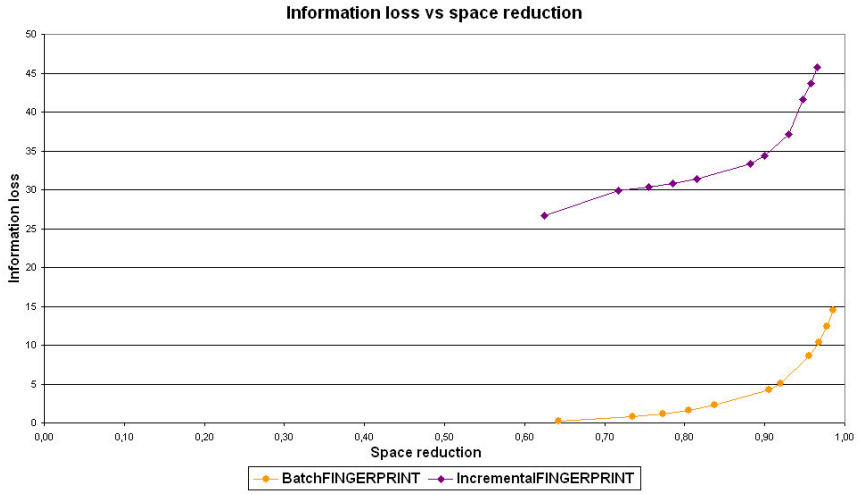
Σχήμα 6.19: Charitable Donation dataset: Επίδραση του κατωφίου δ στην απώλεια πληροφορίας

6.8 Σχετικές εργασίες

Σχετικές με την εργασία μας είναι εργασίες σε παρακολούθηση συστάδων και σε σύνοψη/συμπίεση συστάδων.

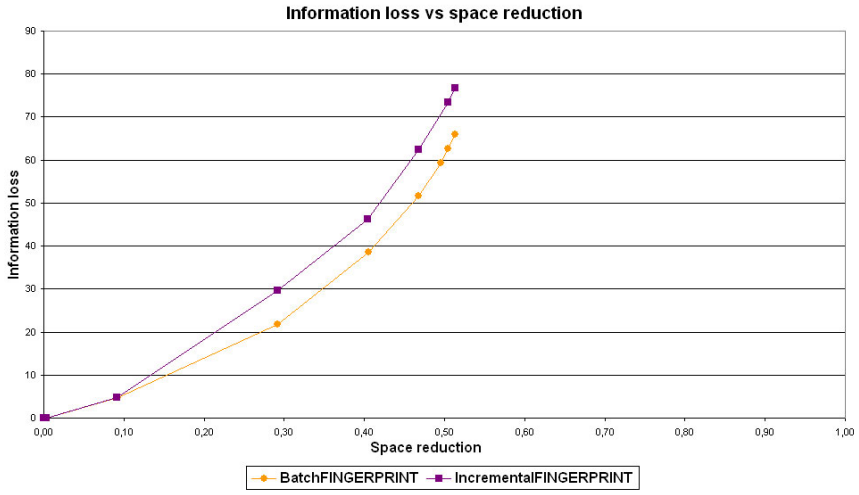


Σχήμα 6.20: ACM H.2.8 dataset: Επίδραση του κατωφίου δ στην απώλεια πληροφορίας

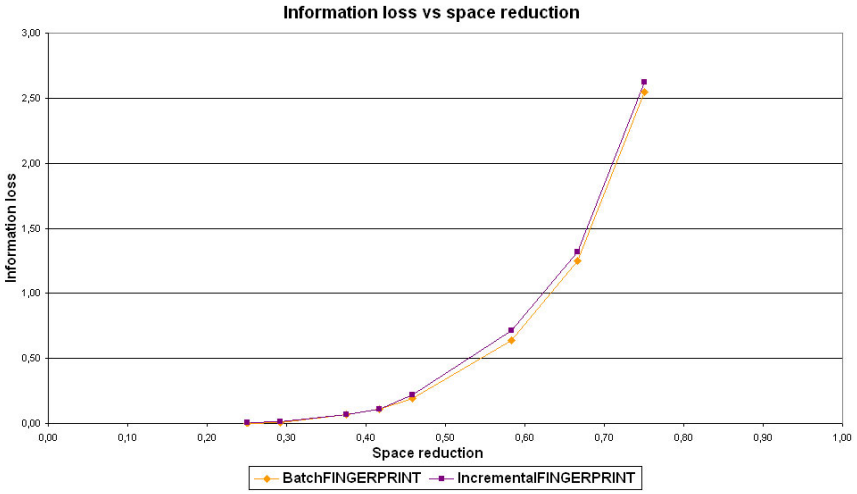


Σχήμα 6.21: Network Intrusion dataset: Συσχέτιση μεταξύ της απώλειας πληροφορίας και του κέρδους συμπαγότητας

Μέθοδοι παρακολούθησης Σχετικές με την εργασία μας είναι μέθοδοι εντοπισμού αλλαγών σε συστάδες (cluster change detection) και μέθοδοι συσταδοποίησης σε χωροχρονικά δεδομένα (spatiotemporal clustering). Επίσης σχετικές είναι και οι μέθοδοι της εξέλιξης θεμάτων (topic evolution).



Σχήμα 6.22: Charitable Donation dataset: Συσχέτιση μεταξύ της απώλειας πληροφορίας και του κέρδους συμπαγότητας



Σχήμα 6.23: ACM H.2.8 dataset: Συσχέτιση μεταξύ της απώλειας πληροφορίας και του κέρδους συμπαγότητας

Εντοπισμός αλλαγών σε συστάδες Το πλαίσιο *FOCUS* [25] συγκρίνει δύο σύνολα δεδομένων με βάση τα μοντέλα εξόρυξης γνώσης που εξάγονται από αυτά. Οι συστάδες αποτελούν μία ειδική κατηγορία μοντέλων που περιγράφονται ως μη επικαλυπτόμενες περιοχές, οι οποίες περιγράφονται μέσω ενός συνόλου γνωρισμάτων (δομική συνιστώσατ) και αντιστοιχούν σε ένα σύνολο δεδομένων (ποσοτική συνιστώσα). Η έμφαση, ωστόσο, στη δουλειά αυτή είναι στη σύγκριση

σύνολων δεδομένων και όχι στην κατανόηση της εξέλιξης των συστάδων.

Στην [47], η *Meila* παρέχει μία περίληψη των σχετικών εργασιών για τη σύγκριση συσταδοποιήσεων που προέρχονται από το ίδιο σύνολο δεδομένων, κάτω από διαφορετικές παραμέτρους εξόρυξης, π.χ., διαφορετικοί αλγόριθμοι ή διαφορετικές παράμετροι για τον ίδιο αλγόριθμο. Οι μέθοδοι αυτές όμως αναφέρονται στη σύγκριση των (διαφορετικών) συσταδοποιήσεων που εξάγονται από το ίδιο σύνολο δεδομένων και έτσι δεν μπορεί να χρησιμοποιηθούν στην γενική περίπτωση συσταδοποιήσεων που εξάγονται από διαφορετικά σύνολα δεδομένων [56].

Συσταδοποίηση χωροχρονικών δεδομένων Στη χωροχρονική συσταδοποίηση, μία συστάδα είναι μία πυκνή περιοχή σε ένα στατικό μετρικό χώρο.

Ο Aggarwal [1] μοντελοποιεί τις συστάδες μέσω συναρτήσεων πυρήνα (kernel functions) και τις αλλαγές μεταξύ συστάδων ως αλλαγές στην πυκνότητα του πυρήνα για κάθε χωρική τοποθεσία της τροχιάς. Η έμφαση δίνεται στον υπολογισμό της αλλαγής της ταχύτητας και στην εύρεση των θέσεων με την μεγαλύτερη ταχύτητα – δηλαδή των επικέντρων. Τρεις διαφορετικοί τύποι αλλαγών μπορούν να θεωρηθούν: i) coagulation δεδομένων που αντιστοιχεί σε συνδεδεμένες περιοχές στα δεδομένα που έχουν πυκνότητα ταχύτητας μεγαλύτερη από ένα κατώφλι ορισμένο από τον χρήστη, ii) διάλυση δεδομένων (data dissolution) που αντιστοιχεί σε συνδεδεμένες περιοχές που η πυκνότητα της ταχύτητας είναι μικρότερη από κάποιο κατώφλι ορισμένο από τον χρήστη και iii) ολίσθηση/μετατόπιση δεδομένων (data shift) σε άλλες περιοχές.

Οι Yang et al [84] εντοπίζουν αλλαγές πάνω σε συστάδες χωρικών δεδομένων. Το πλαίσιό τους υποστηρίζει τέσσερα πρότυπα χωρικών συσχετίσεων (SOAP), συγκεκριμένα *Star*, *Clique*, *Sequence*, και *minLink*, τα οποία χρησιμοποιούνται για να μοντελοποιήσουν τις διαφορετικές αλληλεπιδράσεις μεταξύ των χωρικών αντικειμένων. Τέτοιες μέθοδοι ωστόσο, υποθέτουν ότι ο χώρος των γνωρισμάτων δεν αλλάζει. Κατά συνέπεια, δεν μπορούν να χρησιμοποιηθούν για δυναμικούς χώρους γνωρισμάτων, π.χ., στην εξόρυξη γνώσης από κείμενα, όπου τα γνωρίσματα είναι συνήθως συχνές λέξεις. Επιπλέον, οι ιεραρχικοί αλγόριθμοι συσταδοποίησης δεν μπορούν να χρησιμοποιηθούν από μία τέτοια μέθοδο.

Οι Kalnis et al [39] προτείνουν έναν ειδικό τύπος αλλαγών συστάδων, την κινούμενη συστάδα (moving cluster), τα περιεχόμενα της οποίας μπορεί να αλλάζουν, ενώ η συνάρτηση πυκνότητάς της μπορεί να παραμένει η ίδια κατά τη διάρκεια της ζωής της. Οι συγγραφείς βρίσκουν τις κινούμενες συστάδες με την παρακολούθηση των κοινών εγγραφών μεταξύ των συστάδων που ανήκουν σε διαδοχικές χρονικές στιγμές.

Το *MONIC* πλαίσιο για τον εντοπισμό μεταβολών είναι πιο γενικό, καθώς περιγράφει διαφορετικούς τύπους μεταβολών συστάδων, επιτρέπει επίσης την γήρανση των δεδομένων και δεν απαιτεί η συνάρτηση πιθανότητας μίας κινούμενης συστάδας να είναι σταθερή.

Εξέλιξη θέματος Η ανίχνευση αλλαγών σε συστάδες είναι επίσης σχετική με την εξέλιξη θεμάτων σε ρεύματα κειμένων, όπως εξετάζεται στο [51, 46]. Στις εργασίες αυτές ένα θέμα αποτελείται από μία ετικέτα, η οποία με τη σειρά της αποτελείται από τις πιο επικρατούσες λέξεις μέσα στη συστάδα..

Στην [51], οι Moringa και Yamanishi προτείνουν ένα σύστημα ανάλυσης θεμάτων που εκπληρώνει τρία σημαντικά καθήκοντα: i) τον εντοπισμό της δομής των θεμάτων, δηλαδή την αναγνώριση των τύπων των βασικών θεμάτων που υπάρ-

χουν και πόσο σημαντικά είναι τα θέματα αυτά, ii) τον εντοπισμό της εμφάνισης ενός θέματος και iii) το χαρακτηρισμό ενός θέματος, δηλαδή τον εντοπισμό των χαρακτηριστικών του κάθε βασικού θέματος.

Στην [46], οι Mei και Zhai πρότειναν μία μέθοδο για ανακάλυψη και συμπίεση των εξελικτικών προτύπων των θεμάτων σε ρεύματα κειμένων. Οι συγγραφείς εντοπίζουν τα θέματα σε κάθε περίοδο και χρησιμοποιούν την KL divergence για την εύρεση συνδεδεμένων θεμάτων κατά μήκος του χρόνου, δηλαδή, θέματα με παρόμοιες ετικέτες. Με τον τρόπο αυτό χτίζεται ένας γράφος εξέλιξης θεμάτων ο οποίος μπορεί να χρησιμοποιηθεί για τον εντοπισμό μεταβολών θεμάτων και την ανάλυση του χρόνου ζωής των θεμάτων.

Οι μέθοδοι αυτές είναι εφαρμόσιμες όταν μία ετικέτα συστάδας, που είναι κατανοητή από τον χρήστη, μπορεί να εξαχθεί και να παρακολουθηθεί. Ωστόσο, η απόδοση ετικετών στις συστάδες δεν είναι εφικτή για όλες τις εφαρμογές. Για το λόγο αυτό, το πλαίσιο $MONIC$ εντοπίζει μεταβολές συστάδων αντί για μεταβολές στις ετικέτες των συστάδων.

Μέθοδοι συμπίεσης Η σύνοψη των παρατηρούμενων μεταβολών των συστάδων σε ένα ρεύμα δεδομένων δεν έχει αντιμετωπισθεί πλήρως στη βιβλιογραφία, καθώς ερευνητικά αποτελέσματα στον εντοπισμό των αλλαγών των συστάδων έχουν εμφανισθεί μόλις πολύ πρόσφατα. Σχετική με την παρούσα δουλειά είναι η έρευνα στη σύνοψη ρευμάτων εγγραφών (σε αντίθεση με ρεύματα τιμών) και στην αναγνώριση, χαρακτηρισμό και αναπαράσταση αλλαγών στα πρότυπα.

Η σύνοψη ενός συνόλου συναλλαγών με κατηγορηματικά γνωρίσματα μελετάται από τους Chandola και Kumar [16]. Σε μία από τις μεθόδους τους, οι συγγραφείς ορίζουν συνόψεις συσταδοποιώντας τις συναλλαγές, εξάγοντας τα ζεύγη γνώρισματων/τιμων και χρησιμοποιώντας αυτά τα ζεύγη ως συνόψεις για τις συστάδες. Δεν αντιμετωπίζουν το θέμα της αλλαγής των συστάδων σε ένα ρεύμα δεδομένων, αλλά προτείνουν δύο μετρικές που χαρακτηρίζουν το αποτέλεσμα του αλγορίθμου σύνοψης, το κέρδος συμπαγότητας ("compaction gain") και την απώλεια πληροφορίας ("information loss"). Ακολουθώντας την ίδια λογική, οι μετρικές μας έχουν παρόμοια ονόματα. Ωστόσο, στην εργασία τους, συνοψίζουν τα στατικά δεδομένα με τη χρήση συστάδων, ενώ εμείς συνοψίζουμε εξελισσόμενες συστάδες σε ένα ρεύμα δεδομένων.

Η σύνοψη και η αλλαγή μελετώνται και από τους Ipeirotis et al στην [35], η οποία που ασχολείται με τις αλλαγές των συνόψεων των περιεχομένων βάσεων δεδομένων. Οι συγγραφείς ορίζουν ως σύνοψη περιεχομένου ("content summary") μίας βάσης ένα σύνολο από λέξεις κλειδιά, ζυγισμένες ως προς την σημασία τους μέσα στη βάση. Οι υπηρεσίες meta-search χρησιμοποιούν τέτοιες συνόψεις για την επιλογή των κατάλληλων βάσεων δεδομένων. Η αξιοπιστία μίας τέτοιας σύνοψης μειώνεται καθώς τα περιεχόμενα της βάσης δεδομένων αλλάζουν με τον χρόνο. Έτσι, οι συγγραφείς προτείνουν μεθόδους για την ποσοτικοποίηση των αλλαγών στις συνόψεις. Αυτή η μελέτη αντιμετωπίζει τόσο το θέμα της σύνοψης για εξελισσόμενες βάσεις δεδομένων όσο και της ανακάλυψης των αλλαγών. Ωστόσο, η διατήρηση των συνόψεων σε μία συμπυκνωμένη μορφή είναι πέρα από τους σκοπούς της δουλειάς τους. Από την άλλη, το προτεινόμενο πλαίσιο $FINGERPRINT$ δίνει έμφαση στη σύνοψη των μεταβολών του πληθυσμού.

Η ανακάλυψη και η αναπαράσταση των αλλαγών των συστάδων για ρεύματα κειμένων έχουν μελετηθεί και από τους Mei και Zhai [46]. Οι συγγραφείς εφαρμόζουν μία χαλαρή συσταδοποίηση με mixture models σε κάθε χρονική στιγ-

μή, εξάγουν την αντιπροσωπευτική λίστα λέξεων κλειδιών (το 'θέμα') για κάθε συστάδα, και στη συνέχεια παρακολουθούν την εξέλιξη των λιστών αυτών με το να εντοπίζουν αποκλίσεις μεταξύ των τρεχόντων λέξεων κλειδιών στις λίστες και των προηγούμενων λέξεων. Οι μεταβολές των θεμάτων διατηρούνται σε έναν 'γράφο εξέλιξης θεμάτων', που στη συνέχεια χρησιμοποιείται για τον υπολογισμό του κύκλου ζωής κάθε θέματος (με χρήση Hidden Markov Models).

Οι Agrawal et al [2] προτείνουν το πλαίσιο CluStream για συσταδοποίηση σε δυναμικά ρεύματα δεδομένων. Η διαδικασία της συσταδοποίησης χωρίζεται σε δύο μέρη, το online και το offline: Το online τμήμα αποθηκεύει μία στατιστική σύνοψη (την αποκαλούμενη και *micro-clusters*) περιοδικά και το offline τμήμα την χρησιμοποιεί για τη δημιουργία των πραγματικών συστάδων (τα αποκαλούμενα και *macro-clusters*) σε έναν χρονικό ορίζοντα που καθορίζεται από τον χρήστη. Τα micro-clusters μπορούν να θεωρηθούν ως συνόψεις συστάδων και είναι πράγματι σχεδιασμένα για τη μείωση των απαιτήσεων σε χώρο. Παρ΄ όλα αυτά, το CluStream επικεντρώνεται στο συνδυασμό τους σε συστάδες αντί στη σύνοψή τους. Επίσης, η απώλεια πληροφορίας που προέρχεται από τη σύνοψη δεν μελετάται.

Η σύνοψη και η εξέλιξη συστάδων αντιμετωπίζονται στο CACTUS [24] και το DEMON [26]. Το CACTUS συσταδοποιεί κατηγορηματικά δεδομένα και εξάγει συνόψεις συστάδων, ενώ το DEMON είναι υπεύθυνο για την παρακολούθηση της εξέλιξης των δεδομένων στο χρόνο, εντοπίζοντας και τονίζοντας συστηματικές αλλαγές στα δεδομένα. Με αυτές τις δύο συνιστώσες, είναι δυνατή η μελέτη της εξέλιξης των συνόψεων των δεδομένων, παρόμοια με την εξέλιξη των συνόψεων των κειμένων που μελετήθηκε στο [46]. Ωστόσο, η μοντελοποίηση και η διατήρηση των παρατηρούμενων αλλαγών δεν μελετήθηκε στις εργασίες αυτές.

6.9 Συμπεράσματα

Στο κεφάλαιο αυτό μελετήσαμε το θέμα της παρακολούθησης και του εντοπισμού αλλαγών σε ένα δυναμικό περιβάλλον με βάση τις συστάδες που εξάγονται από το περιβάλλον αυτό.

Αρχικά, προτείναμε το πλαίσιο $MONIC$ για την κατηγοριοποίηση και τον εντοπισμό των μεταβολών των συστάδων. Το πλαίσιο $MONIC$ έχει σχεδιαστεί για αυθαίρετους τύπους συστάδων, γι΄ αυτό έχει υιοθετήσει τον ορισμό των συστάδων ως σύνολα δεδομένων. Εφαρμόσαμε το πλαίσιο $MONIC$ σε μία ενότητα της βιβλιοθήκης ACM και είδαμε πως μπορεί να μας βοηθήσει να αποκτήσουμε κατανόηση σε σχέση με την εξέλιξη του πληθυσμού.

Επεκτείναμε το $MONIC$ στο $MONIC+$ πλαίσιο το οποίο αξιοποιεί επιπλέον και τα ειδικά χαρακτηριστικά κάθε τύπου συστάδας (π.χ., την κατανομή των δεδομένων στην περίπτωση συστάδων που ορίζονται ως κατανομές ή τις τοπολογικές ιδιότητες των συστάδων για συστάδες που ορίζονται πάνω σε κάποιο μετρικό χώρο) στη διαδικασία εντοπισμού των μεταβολών. Τα πειράματά μας δείχνουν πως το πλαίσιο παρέχει χρήσιμη γνώση σε σχέση με την εξέλιξη του πληθυσμού και επίσης ότι, ανάλογα με τον τύπο των θεωρούμενων συστάδων, διαφορετική γνώση επί της εξέλιξης του πληθυσμού προκύπτει.

Οργανώσαμε τις συστάδες και τις μεταβολές τους σε ένα γράφο, τον *Evolution Graph* ο οποίος μοντελοποιεί όλη την ιστορία της εξέλιξης του πληθυσμού. Περιγράψαμε δύο δυνατότητες αξιοποίησης αυτού του γράφου: τις επερωτήσεις πάνω στην εξέλιξη του πληθυσμού και τη μελέτη της ευστάθειας του πληθυσμού με βάση το χρόνο ζωής των εξαγόμενων συστάδων και συσταδοποιήσεων.

Μελετήσαμε επίσης την αποδοτική συμπίεση του *Evolution Graph*καθώς η περίοδος παρακολούθησης αυξάνεται.Για το λόγο αυτό προτείναμε το $FINGERPRINT$, το οποίο συνοψίζει τον *Evolution Graph*έτσι ώστε οι λιγότερο πληροφοριακές συστάδες και μεταβολές μεταξύ συστάδων να απομακρύνονται. Ορίσαμε συναρτήσεις που μετρούν την απώλεια πληροφορίας και το κέρδος συμπαγότητας που προκύπτουν λόγω της συμπίεσης. Υλοποιήσαμε το $FINGERPRINT$ σε δύο εκδόσεις: Την *batch* έκδοση ($Batch\ \ FINGERPRINT$) που συνοψίζει τον *Evolution Graph* ως σύνολο και την *online* έκδοση ($Incremental\ \ FINGERPRINT$) που δημιουργεί τη σύνοψη αυξητικά, καθώς προχωράει η περίοδος της παρατήρησης. Τρέξαμε πειράματα σε τρία πραγματικά σύνολα δεδομένων και είδαμε ότι οι δύο εκδόσεις επιτυγχάνουν παρόμοιο κέρδος συμπαγότητας ενώ η απώλεια πληροφορίας είναι μεγαλύτερη στην *online* έκδοση.

Μέρος αυτής της δουλειάς έχει δημοσιευτεί [74],[73],[75],[72],[56], ενώ μία εκτενής έκδοση έχει υποβληθεί[57].

6.10 Ανοιχτά θέματα

Μία προφανής επέκταση είναι ο εμπλουτισμός των $MONIC$ ($MONIC+$) μέσω της προσθήκης νέων τύπων μεταβολών και δεικτών εντοπισμού αυτών των μεταβολών.

Τα πειράματα έδειξαν πως η παρακολούθηση των συστάδων μπορεί να μας βοηθήσει να κατανοήσουμε την εξέλιξη του πληθυσμού στην πορεία του χρόνου. Ο χαμηλός ρυθμός $passforward$ σε κάποια χρονική στιγμή θα πρέπει να αποτελέσει καμπανάκι για τον τελικό χρήστη και να επιστήσει την προσοχή του στην αντίστοιχη συσταδοποίηση. Μέχρι τώρα, η δουλειά αυτή αφήνεται στο χρήστη, θα ήταν ωστόσο χρήσιμο να διευκολύνουμε τον τελικό χρήστη επισημαίνοντάς του για παράδειγμα τα γνωρίσματα που είναι περισσότερο υπεύθυνα για τις παρατηρούμενες μεταβολές.

Όσον αφορά στο θέμα της συμπίεσης/σύνοψης της εξέλιξης ενός πληθυσμού, μέχρι τώρα, μέσω του πλαισίου $FINGERPRINT$, εστιάσαμε στη συμπίεση/σύνοψη μεταβολών επιβίωσης (survival). Ωστόσο, υπάρχουν και άλλες μεταβολές που μπορεί να υποστεί μία συστάδα όπως διάσπαση ή απορρόφηση. Θα ήταν ενδιαφέρον να δούμε πως η διαδικασία της συμπίεσης μπορεί να εφαρμοστεί και για μεταβολές τέτοιου τύπου, έτσι ώστε τελικά η σύνοψη να αποτελείται αποκλειστικά από εκείνες τις συστάδες που παίζουν σημαντικό ρόλο στην εξέλιξη του πληθυσμού.

Ένα άλλο ενδιαφέρον θέμα είναι η ποιότητα των εξαγόμενων συστάδων και πως αυτή επηρεάζει το αποτέλεσμα της παρακολούθησης και της συμπίεσης. Μέχρι στιγμής, στα πλαίσια των $MONIC$ ($MONIC+$) και $FINGERPRINT$, δεν αξιολογούμε την ποιότητα των συσταδοποιήσεων που λαμβάνουμε από κάποιο αλγόριθμο συσταδοποίησης, παρά μόνο χρησιμοποιούμε τις εξαγώμενες συστάδες ως είσοδο στους διάφορους αλγορίθμους μας. Θα είχε ωστόσο ενδιαφέρον να δούμε πως η ποιότητα των εξαγώμενων συστάδων επηρεάζει το αποτέλεσμα της παρακολούθησης και της συμπίεσης και κατά πόσο αυτή η έννοια της ποιότητας μπορεί να ενσωματωθεί στη διαδικασία της παρακολούθησης και της συμπίεσης.

Κεφάλαιο 7

Συμπεράσματα και Ανοιχτά Θέματα

Στο κεφάλαιο αυτό συνοψίζουμε τα περιεχόμενα της διατριβής και περιγράφουμε κάποια ανοιχτά ερευνητικά θέματα που σχετίζονται με το πρόβλημα της αποτίμησης της ομοιότητας μεταξύ προτύπων.

7.1 Σύνοψη της συνεισφοράς

Εξαιτίας της ευρείας εφαρμογής της Διαδικασίας Ανακάλυψης Γνώσης από τα Δεδομένα και ως αποτέλεσμα της πλημμύρας δεδομένων που παρατηρείται στις μέρες μας, το πλήθος των προτύπων που εξάγονται από ετερογενείς πηγές πληροφορίας (π.χ., επιστήμη, επιχειρήσεις, τηλεπικοινωνίες, Διαδίκτυο) είναι τεράστιο και αρκετά συχνά, μη διαχειρίσιμο από τους χρήστες. Συνεπώς, υπάρχει ανάγκη για αποδοτική διαχείριση προτύπων η οποία περιλαμβάνει θέματα όπως μοντελοποίηση, αποθήκευση και ανάκτηση των εξαγώμενων προτύπων.

Ένα σημαντικό θέμα στο πρόβλημα της διαχείρισης προτύπων είναι αυτό της αποτίμησης της ανομοιότητας μεταξύ προτύπων. Το εν λόγω πρόβλημα αποτελεί ένα δύσκολο πρόβλημα. Καταρχήν, υπάρχει μία πληθώρα από τύπους προτύπων για τους οποίους θα πρέπει να οριστούν τελεστές ανομοιότητας. Επιπλέον, τα πρότυπα μπορεί να ορίζονται τόσο πάνω σε πρωτογενή δεδομένα (τα λεγόμενα απλά πρότυπα) όσο και πάνω σε άλλα πρότυπα (τα λεγόμενα σύνθετα πρότυπα). Επίσης, τα πρότυπα εξ΄ ορισμού διατηρούν μέρος της πληροφορίας που εμπεριέχεται στα πρωτογενή δεδομένα, ωστόσο το ποσοστό αυτής της πληροφορίας εξαρτάται από τις παραμέτρους της εξόρυξης. Όλοι αυτοί οι λόγοι καθιστούν το πρόβλημα της αποτίμησης της ανομοιότητας δύσκολο αλλά και εξαιρετικά ενδιαφέρον.

Στηριζόμαστε στην πληροφορία που υπάρχει στο χώρο των προτύπων για να αποτιμήσουμε την ανομοιότητα μεταξύ προτύπων. Αυτή η πληροφορία μπορεί να αναλυτική, με βάση τα δεδομένα από τα οποία έγινε η εξαγωγή των προτύπων, ή περιγραφική με βάση το νόημα/ έννοια που αναπαριστούν τα πρότυπα.

Στα πλαίσια αυτής της διατριβής αντιμετωπίσαμε διάφορες απόψεις του θέματος της αποτίμησης της ανομοιότητας μεταξύ προτύπων, συγκεκριμένα:

- Προτείναμε το πλαίσιο $PANDA$ για τη σύγκριση τόσο απλών όσο και σύνθετων προτύπων, που ορίζονται πάνω σε πρωτογενή δεδομένα και σε άλλα

πρότυπα, αντίστοιχα. Το $PANDA$ μάλιστα υποστηρίζει τη σύγκριση προτύπων αυθαίρετης πολυπλοκότητας. Στα πλαίσια του $PANDA$, το δύσκολο πρόβλημα της σύγκρισης οποιωνδήποτε σύνθετων προτύπων, αποσυντίθεται στο λιγότερο δύσκολο πρόβλημα της σύγκρισης των επιμέρους προτύπων που τα αποτελούν, και τελικά καταλήγει στο πιο εύκολο πρόβλημα της σύγκρισης απλών προτύπων. Συνεπώς, από τη στιγμή που μπορούμε να εκφράσουμε κάποιο πρόβλημα σύγκρισης στη λογική των απλών - σύνθετων πρότυπων, η σύγκρισή τους μπορεί εύκολα να υποστηριχθεί στα πλαίσια του $PANDA$.

- Διερευνήσαμε το πρόβλημα της αντιστοιχίας της ανομοιότητας στο χώρο των προτύπων με την ανομοιότητα στο χώρο των πρωτογενών δεδομένων για την περίπτωση των συχνών στοιχειοσυνόλων. Συγκεκριμένα, μελετήσαμε την επίδραση του κατωφλίου ελάχιστης υποστήριξης $minSuppport$ και του επιπέδου συμπίεσης του πλέγματος των στοιχειοσυνόλων (συχνά, κλειστά συχνά και μέγιστα συχνά στοιχειοσύνολα) στο αποτέλεσμα της σύγκρισης μεταξύ συνόλων από στοιχειοσύνολα. Τόσο τα θεωρητικά όσο και τα πειραματικά μας αποτελέσματα δείχνουν πως μία τέτοια αντιστοιχία είναι υποκειμενική και εξαρτάται από τις παραμέτρους της εξόρυξης που χρησιμοποιήθηκαν για την εξαγωγή των προτύπων.

- Προτείναμε ένα πλαίσιο για τη σύγκριση μεταξύ δέντρων απόφασης (και συνόλων δεδομένων κατηγοριοποίησης) εκμεταλλευόμενοι την πληροφορία που περιέχουν τα μοντέλα δέντρων απόφασης. Συγκεκριμένα, η προσέγγισή μας αξιοποιεί την πληροφορία που παρέχεται από την τμηματοποίηση που δημιουργεί ένα δέντρο απόφασης πάνω στο χώρο των γνωρισμάτων του προβλήματος κατηγοριοποίησης που περιγράφει. Δείξαμε την χρησιμότητα και την εφαρμοσιμότητα του πλαισίου αυτού στον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ δέντρων απόφασης.

- Τέλος, συγκρίναμε συστάδες και συσταδοποιήσεις και δείξαμε μία ακόμα εφαρμογή της σύγκρισης στην περίπτωση της παρακολούθησης της εξέλιξης ενός πληθυσμού. Συγκεκριμένα, μελετήσαμε πως εξελίσσεται ένας πληθυσμός μελετώντας την εξέλιξη των συστάδων που εξάγονται από τον πληθυσμό αυτό. Για το σκοπό αυτό προτείναμε τα πλαίσια $MONIC$ και $MONIC+$ για την κατηγοριοποίηση και τον εντοπισμό των αλλαγών. Το $MONIC$ είναι ανεξάρτητο από τον εκάστοτε τύπο συστάδων, ενώ το $MONIC+$ λαμβάνει υπόψη και τα ιδιαίτερα χαρακτηριστικά κάθε τύπου συστάδων (ιεραρχικός, διαμεριστικός, με βάση την πυκνότητα). Οργανώσαμε τις συστάδες και τις αλλαγές αυτών σε ένα γράφο, καλούμενο $Evolution\ Graph$, ο οποίος περιέχει όλη την ιστορία των αλλαγών του πληθυσμού. Καθώς όμως η περίοδος παρατήρησης του πληθυσμού αυξάνεται, ο γράφος αυτός διογκώνεται και καθίσταται μη-διαχειρίσιμος από τον τελικό χρήστη. Για το σκοπό αυτό, προτείναμε το πλαίσιο $FINGERPRINT$, το οποίο συνοψίζει το γράφο κατά τρόπο αποδοτικό. Όλα μαζί τα $MONIC$, $MONIC+$,$Evolution\ Graph$και $FINGERPRINT$ προσφέρουν την απαραίτητη υποδομή για την παρακολούθηση και την αξιοποίηση των αλλαγών σε ένα δυναμικό περιβάλλον.

7.2 Ανοιχτά θέματα

Διάφορα ενδιαφέροντα θέματα, μερικά από τα οποία περιγράφονται παρακάτω, παραμένουν ανοιχτά.

- Το πλαίσιο $PANDA$για τη σύγκριση προτύπων αυθαίρετης πολυπλοκότητας παρέχει την απαραίτητη υποδομή για το πρόβλημα της αποτίμησης της ομοιότητας καθώς επιτρέπει στο χρήστη να δημιουργεί τα κατά περίπτωση κατάλληλα μέτρα απόστασης απλά αρχικοποιώντας τα επιμέρους δομικά συστατικά του πλαισίου. Ωστόσο, αυτή η αρχικοποίηση μπορεί να αποδειχτεί δύσκολη καθώς, γενικά, η ανακάλυψη του βέλτιστου μέτρου ομοιότητας για ένα πρόβλημα είναι δύσκολη και μάλλον υποκειμενική. Συνεπώς, θα ήταν εύκολο και ταυτόχρονα ενδιαφέρον να βρει κανείς τις κατάλληλες ρυθμίσεις του $PANDA$για συγκεκριμένες περιπτώσεις προβλημάτων αποτίμησης της ανομοιότητας. Μία λύση θα μπορούσε να είναι η ενσωμάτωση στο $PANDA$ήδη προτεινόμενων λύσεων από τη βιβλιογραφία για συγκεκριμένες περιπτώσεις.

- Στο Κεφάλαιο 4 είδαμε την εξάρτηση της ανομοιότητας από τις παραμέτρους της εξόρυξης και καταλήξαμε ότι η αντιστοίχιση της ανομοιότητας στο χώρο των προτύπων με την ανομοιότητα στο χώρο των πρωτογενών δεδομένων δεν είναι προφανής, αλλά θα πρέπει να λαμβάνει υπόψη τις παραμέτρους της εξόρυξης. Συνεπώς, ένα ανοιχτό θέμα είναι η ανακάλυψη ενός μέτρου ανομοιότητας το οποίο θα είναι πιο εύρωστο στις παραμέτρους εξόρυξης και θα διατηρεί καλύτερα τα χαρακτηριστικά του χώρου των δεδομένων στο χώρο των προτύπων.

- Στο Κεφάλαιο 6 δείξαμε την εφαρμογή των μέτρων ανομοιότητας στην παρακολούθηση προτύπων, για την περίπτωση των συστάδων και των συσταδοποιήσεων. Ένα ανοιχτό θέμα είναι η παρακολούθηση και για άλλους τύπους προτύπων, όπως τα συχνά στοιχειοσύνολα και τα μοντέλα δέντρων απόφασης.

- Η ύπαρξη μέτρων απόστασης μεταξύ προτύπων μας επιτρέπει να εκτελούμε διάφορες εργασίες εξόρυξης γνώσης πάνω σε πρότυπα αντί για πρωτογενή δεδομένα (μετα-εξόρυξη), όπως π.χ., συσταδοποίηση δέντρων απόφασης. Θα ήταν ενδιαφέρον να εφαρμόσουμε τα προτεινόμενα μέτρα ανομοιότητας για τέτοιες εργασίες.

- Οι μέθοδοι που προτάθηκαν στα πλαίσια αυτής της διατριβής αναφέρονται στη σύγκριση προτύπων ίδιου τύπου. Είναι ενδιαφέρον να δούμε πως μπορεί να αποτιμηθεί η ανομοιότητα μεταξύ προτύπων διαφορετικών τύπων, π.χ., ένα δέντρο απόφασης με μία συσταδοποίηση. Μία προφανής λύση θα ήταν η μετατροπή του ενός τύπου στον άλλο και η σύγκριση των προτύπων πάνω σε αυτό τον κοινό τύπο. Πιο κομψές λύσεις όμως θα μπορούσαν να διερευνηθούν. Γενικά μάλιστα η εύρεση ενός μέτρου ανομοιότητας ανεξάρτητα από τον τύπο των προτύπων είναι χρήσιμη τόσο για λόγους έρευνας όσο και για λόγους εφαρμογών.

Bibliography

[1] C. C. Aggarwal. On change diagnosis in evolving data streams. *IEEE Transactions on Knowledge Data Engineering (TKDE)*, 17(5):587–600, 2005.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. Yu. A framework for clustering evolving data streams. In *International Conference on Very Large Data Bases (VLDB)*, pages 81–92. VLDB Endowment, 2003.

[3] C. C. Aggarwal and P. S. Yu. A framework for clustering massive text and categorical data streams. In *SIAM International Conference on Data Mining (SDM)*. SIAM, 2006.

[4] R. Agrawal, T. Imielinksi, and A. Swami. Mining association rules between sets of items ins large databases. In *ACM SIGMOD International Conference on Management of Data(SIGMOD)*, pages 207–216. ACM, 1993.

[5] R. Agrawal, M. Mehta, J. C. Shafer, R. Srikant, A. Arning, and T. Bollinger. The Quest data mining system. In *ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 244–249. AAAI Press, 1996.

[6] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, 2002.

[7] S. Baron, M. Spiliopoulou, and O. Günther. Efficient monitoring of patterns in data mining environments. In *East-European Conference on Advances in Databases and Information Systems (ADBIS)*, pages 253–265. Springer, 2003.

[8] I. Bartolini, P. Ciaccia, I. Ntoutsi, M. Patella, and Y. Theodoridis. A unified and flexible framework for comparing simple and complex patterns. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 496–499. Springer-Verlag New York, Inc., 2004.

[9] I. Bartolini, P. Ciaccia, I. Ntoutsi, M. Patella, and Y. Theodoridis. The PANDA framework for comparing arbitrary complex patterns. Submitted to an international journal, May 2008.

[10] M. A. Bender and M. Farach-Colton. The LCA problem revisited. In *Latin American Symposium on Theoretical Informatics*, pages 88–94. Springer-Verlag, 2000.

[11] C. L. Blake and C. J. Merz. UCI Repository of machine learning databases. http://www.ics.uci.edu/ mlearn/MLRepository.html (valid as of May 2008).

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth International, 1984.

[13] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. volume 30, pages 107–117. Elsevier Science Publishers B. V., 1998.

[14] D. Burdick, M. Calimlim, and J. Gehrke. Mafia: A maximal frequent itemset algorithm for transactional databases. In *International Conference on Data Engineering (ICDE)*, pages 443–452. IEEE Computer Society, 2001.

[15] B. Catania, A. Maffalena, A. Mazza, E. Bertino, and S. Rizzi. A framework for data mining pattern management. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 87–98. Springer-Verlag New York, Inc., 2004.

[16] V. Chandola and V. Kumar. Summarization - compressing data into an informative representation. In *IEEE International Conference on Data Mining (ICDM)*, pages 98–105, Houston, Texas, USA, 2005. IEEE Computer Society.

[17] CWM. Common Warehouse Metamodel (CWM). http://www.omg.org/cwm, (valid as of May 2008).

[18] G. Das and D. Gunopulos. *Time Series Similarity and Indexing*, chapter 11, pages 279–302. Handbook of Massive Data Sets. Kluwer Academic Publishers, 2002.

[19] DIAsDEM. DIAsDEM. http://sourceforge.net/projects/hypknowsys, (valid as of May 2008).

[20] DMG. Predictive Model Markup Language (PMML). http://www.dmg.org/pmml-v3-0.html., (valid as of May 2008).

[21] F. Farnstrom, J. Lewis, and C. Elkan. Scalability for clustering algorithms revisited. *SIGKDD Explorations*, 2(1):51–57, 2000.

[22] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. 1996.

[23] FIMI. Frequent itemsets mining data set repository. http://fimi.cs.helsinki.fi/data/, (valid as of May 2008).

[24] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS: Clustering categorical data using summaries. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 73–83. ACM, 1999.

[25] V. Ganti, J. Gehrke, and R. Ramakrishnan. A framework for measuring changes in data characteristics. In *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 126–137. ACM Press, 1999.

[26] V. Ganti, J. Gehrke, and R. Ramakrishnan. Demon: Mining and monitoring evolving data. In *International Conference on Data Engineering (ICDE)*, pages 439–448. IEEE Computer Society, 2000.

[27] K. Gouda and M. Zaki. Efficiently mining maximal frequent itemsets. In *IEEE International Conference on Data Mining (ICDM)*, pages 163–170. IEEE Computer Society, 2001.

[28] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining Knowledge Discovery (DMKD)*, 15(1):55–86, 2007.

[29] J. Han, Y. Fu, K. Koperski, W. Wang, and O. Zaiane. DMQL: A data mining query language for relational databases. In *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD)*, 1996.

[30] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann Publishers Inc., 2000.

[31] J. Hipp, U. Guntzer, and G. Nakhaeizadeh. Algorithms for association rule mining - a general survey and comparison. *ACM SIGKDD Explorations*, 2(1):58–64, 2000.

[32] IBM. IBM DB2 Intelligent Miner. http://www-306.ibm.com/software/data/iminer, (valid as of May 2008).

[33] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *Communications of the ACM*, 39(11):58–•64, 1996.

[34] T. Imielinski and A. Virmani. MSQL: A Query Language for Database Mining. *Data Mining and Knowledge Discovery*, 3:373–408, 1999.

[35] P. G. Ipeirotis, A. Ntoulas, J. Cho, and L. Gravano. Modeling and managing content changes in text databases. In *International Conference on Data Engineering (ICDE)*, pages 606 – 617. IEEE Computer Society, 2005.

[36] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computer Surveys*, 31:264–323, 1999.

[37] JDMAPI. Java Data Mining API. http://www.jcp.org/jsr/detail/73.prt, (valid as of May 2008).

[38] T. Johnson, L. Lashmanan, and T. Ravmond. The 3W Model and Algebra for Unified Data Mining. In *International Conference on Very Large Data Bases (VLDB)*. VLDB Endowment, 2000.

[39] P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *International Symposium on Spatial and Temporal Databases (SSTD)*, pages 364–381. Springer, 2005.

[40] E. Kotsifakos, I. Ntoutsi, and Y. Theodoridis. Database support for data mining patterns. In *Panhellenic Conference on Informatics (PCI)*. Springer-Verlag, 2005.

[41] H. Kuhn. Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[42] E. Levina and P. Bickel. The Earth Mover's Distance is the Mallow's distance: Some insights from statistics. In *International Conference on Computer Vision (ICCV)*, pages 251–256, 2001.

[43] M. Ley. DBLP The digital bibliography and library project. http://www.informatik.uni-trier.de/ ley/db/ (valid as of May 2008).

[44] T. Li, M. Ogihara, and S. Zhu. Association-based similarity testing and its applications. *Intelligent Data Analysis*, 7:209–232, 2003.

[45] P. Lyman and H. R. Varian. How Much Information? http://www2.sims.berkeley.edu/research/projects/how-much-info/, 2003 (valid as of May 2008).

[46] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 198–207. ACM, 2005.

[47] M. Meila. Comparing clusterings. Technical report, Department of Statistics, University of Washington, 2002.

[48] R. Meo, G. Psaila, and S. Ceri. A New SQL–like Operator for Mining Association Rules. In *International Conference on Very Large Data Bases (VLDB)*, pages 122–133. VLDB Endowment, 1996.

[49] T. Mielikainen. On inverse frequent set mining. In *Workshop on Privacy Preserving Data Mining (PPDM)*, pages 18–23, 2003.

[50] T. Mitchell. *Machine Learning*. Kluwer Academic Publishers, 1997.

[51] S. Morinaga and K. Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 811–816, 2004.

[52] MSSQL. Microsoft SQL Server. http://www.microsoft.com/sql/2005/, (valid as of May 2008).

[53] O. Nasraoui, C. Cardona-Uribe, and C. Rojas-Coronel. Tecno-streams: Tracking evolving clusters in noisy data streams with a scalable immune system learning model. In *IEEE International Conference on Data Mining (ICDM)*, page 235, 2003.

[54] I. Ntoutsi. The notion of similarity in data and pattern space. In *Workshop on Pattern Representation and Management (ParMa) in conjunction with the Int. Conference on Extending Database Technology (EDBT)*. CEUR-WS.org, 2004.

[55] I. Ntoutsi, A. Kalousis, and Y. Theodoridis. A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees. In *SIAM International Conference on Data Mining (SDM)*. SIAM, 2008.

[56] I. Ntoutsi, N. Pelekis, and Y. Theodoridis. Pattern Comparison in Data Mining: a survey. In D. Taniar, editor, *Research and Trends in Data Mining Technologies and Applications (Advances in Data Warehousing and Mining)*, pages 86 – 120. Idea Group Publishing, 2007.

[57] I. Ntoutsi, M. Spiliopoulou, and Y. Theodoridis. Tracing cluster transitions for different cluster types. Submitted to an international journal in February 2008.

[58] I. Ntoutsi and Y. Theodoridis. Current issues in modeling data mining processes and results. In *PANDA Workshop on Pattern-Base Management Systems*, 2003.

[59] I. Ntoutsi and Y. Theodoridis. Measuring and evaluating dissimilarity in data and pattern spaces. In *VLDB International Conference on Very Large Data Bases (VLDB) Phd Workshop*, 2005.

[60] I. Ntoutsi and Y. Theodoridis. The notion of patterns in data mining. Technical Report TR-2007-02, Information Systems Laboratory, Department of Informatics, University of Piraeus, Greece, 2007.

[61] I. Ntoutsi and Y. Theodoridis. Pattern Management. Technical Report TR-2007-01, Information Systems Laboratory, Department of Informatics, University of Piraeus, Greece, 2007.

[62] I. Ntoutsi and Y. Theodoridis. Comparing datasets using frequent itemsets: Dependency on the mining parameters. Submitted to an international conference, May 2008.

[63] ORACLE. Oracle 10g Data Mining Concepts. http://download-uk.oracle.com/docs/cd/B19306-01/datamine.102/b14339/toc.htm, (valid as of May 2008).

[64] PANDA. The PANDA Project. http://dke.cti.gr/PANDA/, (valid as of May 2008).

[65] PANDA. PAtterns for Next generation DAtabase systems - IST project 2001–2003, (valid as of May 2008).

[66] S. Parthasarathy and M. Ogihara. Clustering distributed homogeneous datasets. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 566–574. Springer, 2000.

[67] I. Pekerskaya, J. Pei, and K. Wang. Mining changing regions from access-constrained snapshots: A cluster-embedded decision tree approach. *Intelligent Information Systems (Special Issue on Mining Spatio-Temporal Data)*, 27(3):215–242, 2006.

[68] riskglossary.com. Kurtosis. http://www.riskglossary.com/link/kurtosis.htm, (valid as of May 2008).

[69] riskglossary.com. skewness. http://www.riskglossary.com/link/skewness.htm, (valid as of May 2008).

[70] S. Rizzi, E. Bertino, B. Catania, M. Golfarelli, M. Halkidi, M. Terrovitis, P. Vassiliadis, M. Vazirgiannis, and E. Vrachnos. Towards a Logical Model for Patterns. In *ER*, pages 77–90, 2003.

[71] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision*, pages 56–66. IEEE Computer Society, 1998.

[72] M. Spiliopoulou, I. Ntoutsi, and Y. Theodoridis. Tracing cluster transitions for different cluster types. Technical Report TR-2007-03, Information Systems Laboratory, Department of Informatics, University of Piraeus, Greece, 2007.

[73] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. The MONIC framework for cluster transition detection. In *Hellenic Data Management Symposium (HDMS)*, 2006.

[74] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. MONIC: modeling and monitoring cluster transitions. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 706–711. ACM Press, 2006.

[75] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. Tracing cluster transitions for different cluster types. In *Workshop on Data Mining and Knowledge Discovery (ADMKD) in conjunction with the East-European Conference on Advances in Databases and Information Systems (ADBIS)*, 2007.

[76] M. Spiliopoulou, Y. Theodoridis, and I. Ntoutsi. Mining the Volatile Web - tutorial. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2005.

[77] SQL/MM. ISO SQL/MM Part 6. http://www.sql-99.org/SC32/WG4/Progression-Documents/FCD/fcd-datamining-2001-05.pdf, (valid as of May 2008).

[78] P. Turney. Technical note: Bias and the quantification of stability. *Machine Learning*, 20:23–33, 1995.

[79] J. R. Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM*, 23(1):83–97, 1976.

[80] H. Wang and J. Pei. A random method for quantifying changing distributions in data streams. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 684–691. Springer, 2005.

[81] X. Wu, C. Zhang, and S. Zhang. Database classification for multi-database mining. *Information Systems*, 30(1):71–88, 2005.

[82] D. Xin, J. Han, X. Yan, and H. Cheng. Mining compressed frequent-pattern sets. In *International Conference on Very Large Data Bases (VLDB)*, pages 709–720. VLDB Endowment, 2005.

[83] Yahoo. Italian MIB stock. http://it.finance.yahoo.com/, (valid as of May 2008).

[84] H. Yang, S. Parthasarathy, and S. Mehta. A generalized framework for mining spatio-temporal patterns in scientific data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 716–721. ACM Press, 2005.

[85] M. Zaki and C.-J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(4):462–478, 2005.

[86] K. Zhang and D. Shasha. Fast algorithms for the editing distance between trees and related problems. *SIAM Journal of Computation*, 18(6):1245–1262, 1989.